



**Michigan Educational Assessment Program**

**Technical Report**

**2012-2013**

**Michigan Department of Education  
Bureau of Assessment and Accountability  
And Measurement Incorporated**

**Compiled by  
Dong Gi Seo**

## TABLE OF CONTENTS

<b>Introduction and Overview of Technical Report</b> .....	1
<b>Chapter 1: BACKGROUND OF THE MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM (MEAP)</b> .....	2
<b>Chapter 2: Test Development</b> .....	6
2.1. Test Specifications.....	6
2.1.1. Item Writer Training .....	6
2.1.2. Item Development.....	6
2.1.3. Item Review.....	8
2.1.4. Field Testing.....	8
2.1.5. Data Review.....	8
2.1.6. Operational Test Construction.....	8
2.2. Released Items/ Item Descriptor Reports.....	9
2.2.1. Test Structures for 2012 MEAP Content Tests.....	9
2.3. Review of Field Test Items Provided by Development Contractor.....	12
2.3.1. Tabulations of Item Characteristics.....	12
2.3.2. Item Specifications.....	12
2.3.3. Item Statistics.....	12
2.3.4. Differential Item Functioning.....	12
2.3.5. Data Review.....	13
2.4. Pre-Field-Test Item Review.....	13
2.4.1. Contractor Review.....	13
2.4.2. BAA Review.....	14
2.5. Field Testing.....	14
2.5.1. Field Testing Design.....	14
2.5.2. Field Testing Sampling.....	14
2.6. Data Review.....	16
2.6.1. Data.....	16
2.6.2. Statistics Prepared for Review Committees .....	17
2.6.3. Data Reviews.....	21
2.6.3.1. Bias/Sensitivity and Content Committee Review.....	21
2.6.4. Item Revision Procedures.....	21
2.7. Item Banking Procedure.....	25
2.8. Construction of Operational Test Forms.....	25
2.8.1. Design of Test Forms.....	25
2.8.1.1. Review the Assessment Blueprints for the Operational Assessments.....	25
2.8.2. Item Selection.....	28
2.8.2.1. Select Assessment Items to Meet the Assessment Blueprints .....	29
2.8.2.2. Assess the Statistical Characteristics of the Selected Assessment Items.....	29
2.8.2.3. Review and Approve Test Forms.....	30
2.9. Accommodated Test Forms.....	31

2.9.1. Special Order Accommodated Testing Materials.....	31
2.9.2. Braille.....	31
2.9.3. Large Print.....	31
2.9.4. Oral Administration for Mathematics.....	32
2.9.5. Bilingual Tests.....	32
<b>Chapter 3: Overview of Test Administration.....</b>	<b>33</b>
3.1. Test Administration.....	33
3.2. Materials Return.....	34
<b>Chapter 4: Technical Analyses of Post-Administration Processing.....</b>	<b>38</b>
4.1. Scanning Accuracy and Reliability.....	38
4.2. Multiple-Choice Scoring Accuracy.....	41
4.3. Erasure Analyses.....	42
4.4. Results of Constructed Response Scoring Procedures.....	42
4.4.1. Rangefinding and Rubric Review.....	43
4.4.2. Rater Selection.....	45
4.4.3. Rater Training.....	46
4.5. Rater Statistics and Analyses.....	48
4.5.1. Calibration.....	48
4.5.2. Rater Monitoring and Retraining .....	49
4.5.3. Rater Dismissal .....	50
4.5.4. Score Resolution.....	50
4.5.5. Inter-Rater Reliability Results.....	50
4.5.6. Rater Validity Checks.....	50
<b>Chapter 5: MEAP Reports.....</b>	<b>51</b>
5.1. Description of Scores.....	51
5.1.1. Scale Score.....	51
5.1.2. Raw Score.....	51
5.1.3. Performance Level .....	51
5.1.4. Mini-Categories.....	52
5.1.5. Performance Level Change.....	52
5.2. Scores Reported.....	53
5.3. Appropriate Score Uses.....	54
5.3.1. Individual Students.....	54
5.3.2. Groups of Students.....	54
5.3.3. Item Statistics.....	55
5.3.4. Frequency Distributions.....	57
<b>Chapter 6: Performance Standard.....</b>	<b>58</b>
6.1. Development of Standard Setting Performance Level Descriptors.....	58
6.2. Standard Setting.....	62
6.3. Revised Standards for Writing.....	62
6.3.1. Standard Setting Methodology.....	63

6.3.2. Selection of Panelists.....	66
6.3.3. Standard Setting.....	67
6.3.3.1. Round 1.....	67
6.3.3.2. Round 2.....	69
6.3.3.3. Round 3.....	71
6.3.4.4. Final Standard Determinations.....	73
6.4. Revised Proficiency Level Cut Scores.....	73
<b>Chapter 7: Scaling.....</b>	<b>76</b>
7.1. Summary Statistics and Distributions from Application of Measurement Models.....	77
7.1.1. Summary Classical Test Theory Analyses by Form.....	77
7.1.2. Item Response Theory Analyses by Form and for Overall Grade-Level Scales.....	77
7.1.2.1. Step-by-Step Description of Procedures Used to Calibrate Student Responses, Using Item Response Theory.....	77
7.1.2.2. Summary Post-Calibration Score Distributions.....	78
7.1.2.3. Summary Statistics on Item Parameter Distributions and Fit Statistics.....	89
7.1.2.4. Test Information/Standard Error Curves.....	96
7.1.2.5. Summary of Model fit Analyses.....	106
7.2. Scale Scores.....	106
7.2.1. Description of the MEAP Scale.....	106
7.2.2. Identification of the Scale, Transformation of IRT Results to MEAP Scale.....	106
7.2.3. Scale Score Interpretations and Limitations.....	107
7.2.4. Upper and Lower End Scaling.....	109
<b>Chapter 8: Equating.....</b>	<b>110</b>
8.1. Rationale.....	110
8.2. Pre-equating.....	110
8.2.1. Test Construction and Review.....	110
8.2.2. Field-Test Items.....	110
8.2.3. Within-Grade Equating.....	111
8.2.3.1. Description of Procedures Used to Horizontally Equate Scores from Various Forms at the Same Grade Level.....	111
8.2.3.2. Item and Test Statistics on the Equated Metric, Test Information Standard Error Curves, and Ability Distributions.....	111
8.3. Vertical Equating.....	111
8.4. Ability Estimation .....	111
8.5. Development Procedures for Future Forms.....	113
8.5.1. Equating Field-Test Items.....	113
8.5.2. Item Pool Maintenance.....	113
<b>Chapter 9: Reliability.....</b>	<b>114</b>
9.1. Internal Consistency, Empirical IRT Reliability Estimates, and Conditional Standard Error of Measurement.....	114
9.1.1. Internal Consistency.....	114

9.1.2. Empirical IRT Reliability.....	116
9.1.3. Conditional Standard Error of Measurement.....	117
9.1.4. Use of the Standard Error of Measurement.....	118
9.2. Alternative Forms Reliability Estimates.....	118
9.3. Score Reliability for the Written Composition and the Constructed-Response Items.....	118
9.3.1. Reader Agreement.....	118
9.3.2. Score Appeals.....	119
9.4. Estimates of Classification Accuracy.....	119
9.4.1. Statewide Classification Accuracy.....	119
<b>Chapter 10: Validity.....</b>	<b>121</b>
10.1. Content and Curricular Validity.....	121
10.1.1. Relation to Statewide Content Standards.....	121
10.1.1.1. MEAP Alignment Studies.....	122
10.1.2. Educator Input.....	124
10.1.3. Test Developer Input.....	124
10.1.4. Evidence of Content Validity.....	124
10.2. Criterion and Construct Validity.....	125
10.2.1. Criterion Validity.....	125
10.2.2. Construct Validity.....	126
10.3. Validity Evidence for Different Student Populations.....	130
10.3.1. Differential Item Functioning (DIF) Analyses.....	130
10.3.1.1. Editorial Bias Review.....	131
10.3.1.2. Statistical DIF Analyses.....	132
10.3.1.3. DIF Statistics for Each Item.....	132
10.3.2. Performance of Different Student Populations.....	132
10.4. Validity Evidence for Accommodation from (Person-fit Analysis).....	133
10.5. Validity Evidence for Mode Comparability (Online vs. Paper-Pencil Tests) .....	134
10.6. MEAP Math Rescoring Issues .....	134
<b>Chapter 11: Accountability Uses of Assessment Data .....</b>	<b>137</b>
11.1. Legislative Grounding .....	137
11.2. Procedures for Using Assessment Data for Accountability .....	138
11.3. Results of Accountability Analyses .....	142
<b>References .....</b>	<b>143</b>
<b>List of Appendices .....</b>	<b>144</b>

## INTRODUCTION AND OVERVIEW OF TECHNICAL REPORT

This technical report is designed to provide information to Michigan coordinators, educators and interested citizens about the development procedures and technical attributes of the state-mandated Michigan Educational Assessment Program (MEAP). This report does not include all the information available regarding the assessment program in Michigan. Additional information is available on the Michigan Department of Education (MDE), Office of Educational Assessment & Accountability (OEAA) website.

This report outlines the necessary steps and presents supporting documentation so that educators can improve teaching and learning through the use of assessment results. The information in this report may be used to monitor school and individual student improvement over time. Additionally, this report outlines current “state of the art” technical characteristics of assessment and should be a useful resource for educators trying to explain to parents, teachers, school boards and the public alike the different ways in which assessment information is important.

This technical report includes 10 chapters:

- Chapter 1 gives the general background of the MEAP assessment program, the appropriate uses for the scores and reports, and the organizations and groups involved in the development and administration of the program.
- Chapter 2 describes details of the test specifications and test blueprints, as well as the full cycle of the test development process including item writing, pre-field-test item review, field testing, post-field-testing item review, item banking, and the construction of operational and accommodated test forms.
- Chapter 3 provides an overview of test administration. Activities involved in the preparation for test administration, test administration process, test materials return, measures of test security, and the test accommodations for students with disabilities and students in ELL are presented in this chapter.
- Chapter 4 presents the technical analyses of post-administration processing. Scanning accuracy and reliability, as well as the rater validity and reliability of scoring constructed responses items are discussed in detail.
- Chapter 5 describes the score reporting of the assessment. It includes the descriptions of scale score, raw score and proficiency levels, the type of score reports, and the appropriate score uses.
- Chapter 6 gives a detailed report of the development of performance level descriptors (PLDs), as well as the procedures, implementation, and results of performance standard setting process.
- Chapters 7 through 10 describe the psychometric characteristics of the MEAP assessments. Step-by-step description of procedures used to calibrate student responses using item response theory, development of MEAP scale score, and the rationale and procedures of equating (including 3PL equating for writing) are described in Chapter 7 and 8. Alpha reliability, empirical IRT reliability, reader agreement, estimates of statewide classification accuracy of MEAP, validation of content validity, construct validity, and the performance of different student populations are described in Chapter 9 and Chapter 10. Results of Accountability is described in Chapter 11.

There are also extensive appendices to this report. These are listed at the end of the main report text, and are made available separately due to their size.

## **CHAPTER 1: BACKGROUND OF THE MICHIGAN EDUCATIONAL ASSESSMENT PROGRAM (MEAP)**

Michigan's educational system consists of 57 Intermediate School Districts with 550 local school districts and public school academies. Private Schools and home schooled students also have the option to participate in the Michigan Educational Assessment Program. Altogether, schools serve over 1.5 million students.

The primary function of the Office of Standards and Assessment (OSA) is to establish, develop and conduct an assessment that fairly and accurately reflect Michigan's adopted content standards. The OSA staff directs the implementation and administration of MEAP. In addition to planning, scheduling, and directing all assessment activities, the staff is extensively involved in item construction, item reviews for content and bias, test composition, security, and quality control procedures.

The Michigan Educational Assessment Program (MEAP) is a statewide assessment program first initiated by the State Board of Education in 1969. MEAP is a criterion-referenced assessment for students in grades three through nine in the following content areas:

<b>Grade</b>	<b>Math</b>	<b>Reading</b>	<b>Writing</b>	<b>Science</b>	<b>Social Studies</b>
<b>3</b>	X	X			
<b>4</b>	X	X	X		
<b>5</b>	X	X		X	
<b>6</b>	X	X			X
<b>7</b>	X	X	X		
<b>8</b>	X	X		X	
<b>9</b>					X

MEAP is based on Grade Level Content Expectations (GLCE) and is currently administered in the fall to assess prior year standards. Based on their IEP, students with disabilities take part in MEAP, MEAP-Access, or one of three MI-Access assessments: Participation, Supported Independence, and Functional Independence.

Michigan uses MEAP achievement data to provide report cards to districts and schools. The report cards are made public and used in a variety of ways to hold schools accountable, improve schools, and help parents make decisions about their children's education.

MEAP scores are divided into four performance levels: Not Proficient, Partially Proficient, Proficient, and Advanced. Students who score in either Proficient or Advanced levels are considered to be "proficient" with those content expectations. Those who place in the Not Proficient or Partially Proficient levels are deemed to be "not proficient." Achievement levels are publically reported for the following subgroups:

- Major racial/ethnic subgroups
  - Black or African American

- American Indian or Alaska Native
- Asian
- Native Hawaiian, or other Pacific Islander
- Hispanic of Any Race
- Two or More Races
- White
- Gender
- Students with disabilities
- Limited English proficient
- Economically disadvantaged

MEAP assessments comply with the “*No Child Left Behind Act of 2001*” (NCLB).

### **Use of Scores and Guide To Reports**

Following administration of MEAP assessments, reports and data files are provided to help educators understand and use assessment results. The reports provide educators, parents, and the public with an understanding of the status and progress of Michigan students.

Properly used, MEAP assessment results can be used to:

- measure academic achievement as compared with expectations, and provide a basis for measurement of improvement over time
- determine whether improvement programs and policies are having the desired effect
- focus academic help where it is needed.

A Guide to Reports was developed to assist educators in understanding and using MEAP results. Both individual and aggregate level reports are included in the Guide. The Guide to Reports can be found at [www.michigan.gov/meap](http://www.michigan.gov/meap) and include:

#### **Individual Reports:**

- Individual Student Report: The Individual Student Report provides detailed information on individual student achievement and includes scale score, performance level, possible points, and earned points.
- Parent Report: The Parent Report summarizes individual student achievement and performance level change information (if available).

- **Class Roster:** The Class Roster provides detailed information on student achievement and is sorted by class and group codes (if provided).
- **Student Record Label:** The Student label summarizes student achievement and is provided for students' permanent records

#### **Aggregate Reports:**

- **Comprehensive Report:** The Comprehensive Report provides a summary of the number of students tested, the mean scale score, and performance level information for a district or ISD.
- **Demographic Report:** The Demographic Report summarizes the total number of students tested, the mean scale score, and performance level for each demographic subgroup containing at least ten students.
- **Item Analysis Report:** The Item Analysis identifies and describes each GLCE assessed and provides individual item statistics, including the percentage of students selecting each response.
- **Summary Report:** This report summarizes student achievement for all content areas including mean scale score and performance level information, as well as year to year comparisons. The School Summary also provides student score distributions for each content area.

#### **MEAP Contractor and Subcontractors**

Measurement Incorporated is the development contractor for MEAP. They arrange for test printing, shipping, scoring, and reporting. Cheeney Media Concepts <sup>2</sup> and the American Printing House for the Blind, Inc. serve as subcontractors for the production of accommodated materials. SourceHOV serves as a subcontractor for the printing of student assessment results and reports. Assessment Evaluation Services (AES) provides psychometric support and validation.

#### **Involvement of Many Stakeholders**

The development of MEAP is a meticulous process involving hundreds of Michigan administrators, teachers, and curriculum experts. The OSA actively seeks input and feedback in the development and implementation of assessment and accountability systems to further the educational goal of improving what students know and can do in relation to the state grade level content expectations.

The State Board of Education provides leadership and general supervision over all public education, including adult education and instructional programs in state institutions, with the exception of higher education institutions granting baccalaureate degrees. The State Board of Education serves as the general planning and coordinating body for all public education, including higher education, and advises the legislature concerning the financial requirements of public education.

The Technical Advisory Committee (TAC) was first established in 1993 to assist the MDE in developing a high school proficiency assessment as a requirement for high school graduation as required by PA 118 of 1991. The TAC is made up of individuals from Michigan and across the nation who are recognized experts in developing or reviewing high stakes assessment programs.

The TAC advises and assists the OSA to ensure MEAP assessments are developed in keeping with technical guidelines that meet national standards and independently monitor all assessment development and implementation processes, including information gathered in field tests and review of item development. The TAC may make recommendations for revisions in design, administration, scoring, processing, or use in the assessment.

## **Chapter 2: Test Development**

### **2.1. Test Specifications**

As noted in the previous chapter, all MEAP tests are based on the grade-level content expectations (GLCEs). A general description of development activities applying to all tests is provided below, followed by subject-specific descriptions.

BAA test development staff, contractors, and Michigan educators worked together to develop the tests. The test development cycle included the following steps:

- Item Writer Training
- Item Development
- Item Review
- Field Testing
- Field Test Item Review
- Operational Test Construction

#### **2.1.1. Item Writer Training**

Once item specifications are finalized, experienced contractors use customized materials to train item writers to produce items specifically for MEAP. Item Writer Training can last anywhere from three to five days and is conducted by contractor staff with BAA test development staff oversight. The actual writing of items, including writing items receiving feedback from contractor staff, takes anywhere from 4 to 8 weeks. All item writers are Michigan educators who have curriculum and instruction expertise and who have been recommended by their administrators. They also possess relevant degrees and experience, and many have previous experience in MEAP-specific item writing.

#### **2.1.2. Item Development**

Michigan item writers draft test items in accordance with specifications approved by BAA test development staff. Contractor staff review items internally and then share with MEAP staff for an additional review. This internal review consists of meeting the following criteria:

Skill:

- Item measures one skill level.
- Item measures skill in manner consistent with specification.
- Item uses appropriate (realistic) level of skill.
- Item makes clear the skill to be employed.

Content:

- Item measures one benchmark.
- Item measures benchmark in manner consistent with specification.
- Item taps appropriate (important) aspect of content associated with benchmark.
- Item makes clear the benchmark or problem to be solved.

Relevance:

- Item calls for a realistic application of process to content.
- Item is not contrived.
- Item is appropriate for the grade level to be tested.
- Item groups reflect instructional emphasis.

Accuracy:

- Item is factually accurate.
- Item contains only one correct or best response.
- If item pertains to disputed content, context for correct answer is clearly defined (e.g., "According to... the correct solution is...").
- Item is unambiguously worded.

Format:

- Item contains no extraneous material except as required by the benchmark.
- Vocabulary is grade-appropriate and clear.
- Item contains no errors of grammar, spelling, or mechanics.
- Item responses are parallel and related to the stem.
- Item responses are independent.
- Item contains no clues or irrelevant distracters.
- Directions for responding to a constructed response (CR) item are clear.
- CR item and rubric match.
- CR rubric is clear and easy to apply.
- Item is clearly and conveniently placed on the page.
- Item contains adequate white space for calculations as needed.
- Physical arrangement of item is consistent with benchmark or common practice (e.g., horizontal vs. vertical addition and subtraction, slash vs. horizontal fraction bar, notation, symbols, etc.).
- Keys for sets of multiple choice (MC) items are balanced (i.e., equal numbers of A's, B's, C's, and D's).

Bias:

- Item is free of race and gender stereotypes.
- Item contains no material known or suspected to give advantage to any group.
- Item is free of insensitive language.
- Item sets that identify race or gender either directly or indirectly are balanced with reference to race and gender.
- Item content and format are accessible to students with disabilities.
- Item content and format are accessible to students with limited English proficiency.

### 2.1.3. Item Review

After the internal reviews take place, all MEAP items are reviewed by Michigan educators and/or members of a Michigan community that serve on one of two committees: Content Advisory Committee (CAC) or/and the Bias and Sensitivity Review Committee (BSC). Contractor staff trains the CAC and BSC **participants** and facilitate the committee meetings. All items are typically first reviewed by the BSC and then sent to the CAC.

An item rejected by the BSC may or may not get passed on to the CAC for review. Each review is led by experienced contractor staff with MEAP staff in attendance, using prescribed guidelines and forms to indicate the final status of each item:

- **Accept:** Each of the eight category conditions (importance, thematic, grammar, clarity, accuracy, validity, sound measurement, grade-appropriate) have been met or exceeded and the item appears suitable for field testing.
- **Modify:** One or more of the category conditions have not been met or the item needs minor changes to make it acceptable. Reviewers provide recommendations on changes to be made to the item that will make the item suitable for field testing.
- **Reject:** Several category conditions have not been met, or are suspect, or need radical changes to make the item acceptable. In such cases, the item may be vague or ambiguous, inappropriate, or not clearly related to the text or to the standard. Without severe modifications it is unlikely to be salvaged. Reviewers provide comments to explain why the item should be rejected.

### 2.1.4. Field Testing

Items that have passed bias/sensitivity and content review are then eligible for field testing. MEAP field testing is carried out by embedding items in operational test forms.

### 2.1.5 Data Review

After field testing, contractor staff analyzes results and present them to the same groups listed under Item Review above (BSC and CAC). During these review committees, participants review the items with field test statistics. CAC members review performance data (percent correct, response distribution, raw score distribution, point biserial correlation with total score), while BSC members review data that would indicate differential item functioning (percent correct by group, chi-square and other statistics). Members have the option to accept or reject the item. Once items and their field test results have been presented to the BAA and the BAA has accepted them, they go into an eligible bank of items from which future operational tests may be constructed. Results from the Data Review meetings that occurred during the 2012-2013 administration cycle are discussed later in this chapter.

### 2.1.6 Operational Test Construction

Once items have survived all reviews and field testing, they are placed in an item bank and are eligible for operational use. Contractor staff then select items from the bank that meet the test specifications (i.e., blueprint and psychometric specifications). They present these items to BAA test development staff in several stages, first as one item per page, then as a draft formatted test, and finally as a final formatted

test booklet. In this final stage, a spreadsheet accompanies the test, showing the item code, key, content standard, the statistical/psychometric information for each item, and the projected total test statistics.

## **2.2. Released Items/ Item Descriptor Reports**

On an annual basis the BAA reviews current item inventory and reviews the ability to release any MEAP assessment items. The BAA is able to provide an item descriptor report for each content area which does aide in the review and understanding of the data provided to the schools, parents, and the public.

The Item Descriptor Report contains an item descriptor for each scored item on the assessment. Content staff and item writers work to create the descriptor, or item rationale, for every item. The descriptor describes the stem of each item, and provides reasoning to explain why each possible distractor would not have been the correct answer. These booklets can be tied directly back to the reports provided by using the "descriptor position" identified on the report and can be used to review assessment data in a more thorough manner.

Grades 3-5 Item Descriptors, Released Items, Scoring Guides

[http://www.michigan.gov/mde/0,4615,7-140-22709\\_31168-281205--,00.html](http://www.michigan.gov/mde/0,4615,7-140-22709_31168-281205--,00.html)

Grades 6-8 Item Descriptors, Released Items, Scoring Guides

[http://www.michigan.gov/mde/0,4615,7-140-22709\\_31168-281206--,00.html](http://www.michigan.gov/mde/0,4615,7-140-22709_31168-281206--,00.html)

Grade 9 Item Descriptors, Released Items, Scoring Guides

[http://www.michigan.gov/mde/0,4615,7-140-22709\\_31168-281209--,00.html](http://www.michigan.gov/mde/0,4615,7-140-22709_31168-281209--,00.html)

### **2.2.1 Test Structures for 2012 MEAP Content Tests**

The 2012 MEAP assessment contains five content area tests: reading, writing, mathematics, science, and social studies. Reading and Mathematics tests span grades 3 to 8. For 2012, Writing was only administered in grades 4 and 7. Science was tested in grades 5 and 8, and Social Studies was tested in grades 6 and 9. The test structures are summarized in this section.

#### ***ELA***

The MEAP English Language Arts Assessment is based on the Michigan GLCEs, which have been categorized as Core or Not Assessed at the State Level (NASL). ELA has been broken into two tests, Reading and Writing. Reading is assessed at grades 3-8 and Writing is assessed at grades 4 and 7. Each form of the Reading assessment includes a pair of related texts along with independent items and cross-text items. There are also independent texts. Texts are assessed using both multiple choice items and 3-point short answer items. The Writing assessment includes student writing samples, each with a set of multiple choice items. Students write a response to a student writing sample and to both a Narrative and Informational writing prompt. The Core designations of the GLCEs are available in the English Language Arts Assessable GLCEs document. The Reading assessment consists of three parts. The field test portion generally makes up Part 3 of each form. The Writing assessment consists of five parts, with

the field test making up Parts 1 and 2. For both Reading and Writing assessments, one core form and 5 FT forms were constructed. The core and FT forms are identical across grades. The test structures for ELA forms are summarized in Tables 2.2.1.1 through 2.2.1.5.

**Table 2.2.1.1**  
**Test Structure for Fall 2012 Grades 3-8 ELA Core Tests**

<b>Grade</b>	<b># MC Items</b>	<b># CR Items</b>	<b># Total Core</b>
3-8	30	1	31

**Table 2.2.1.2**  
**Test Structure for Fall 2012 Grades 4 and 7 Writing Core Tests**

<b>Grade</b>	<b># MC Items</b>	<b># CR Items</b>	<b># Total Core</b>
4, 7	16	3	19

**Table 2.2.1.3**  
**Test Structure for Fall 2012 Grades 3-8 Reading FT Tests**

<b>Grade</b>	<b># MC Items</b>	<b># CR Items</b>	<b># Total FT per Grade</b>
3-8	22	2	24

**Table 2.2.1.4**  
**Test Structure for Fall 2012 Grades 4 and 7 Writing FT Tests**

<b>Grade</b>	<b># MC Items</b>	<b># CR Items</b>	<b># Total FT per Grade</b>
4, 7	8	1	9

### ***Mathematics***

The MEAP Mathematics Assessment is based on Michigan GLCEs, which are categorized as Core, Extended Core and, Not Assessed at the State Level (NASL). Each mathematics form includes a common set of two MC items per Core GLCE, or one per Extended Core. Each form consists of two parts. Part 1 is the non-calculator part of the form with Part 2 containing the calculator portion. All Grade 3 items are non-calculator items. For 2012, the core forms for Mathematics consisted of core and field test items. In addition, 5 FT forms were constructed. The test structure for grades 3 through 8 mathematics assessment forms is summarized in Table 2.2.1.5.

**Table 2.2.1.5**  
**Test Structure for Fall 2012 Grades 3 through 8 Mathematics Core Tests**

Grade	# Core			# Total Core	# FT		# Field Test Per Form
	MC	CR	Extended		MC	CR	
3	32	0	13	53	8	0	8
4	40	0	13	59	9	0	9
5	32	0	20	54	10	0	10
6	38	0	14	60	9	0	9
7	38	0	15	62	10	0	10
8	40	0	11	49	8	0	8

### *Science*

For Science tests, each form consists of core and FT items. One core form and 5 FT forms were constructed for elementary school and middle school. The test structure for science tests is summarized in Table 2.2.1.6.

**Table 2.2.1.6**  
**Test Structure for the Fall 2012 Science Tests**

Grade	# MC	# CR	# Total Core	# Field Test Item
5	48	0	32	12
8	53	0	33	12

### *Social Studies*

For the Social Studies tests, one core form and 5 FT forms was constructed for elementary school and middle school. The Grade 6 FT items were developed to the new Grade Level Content Expectations (GLCE) for grades 3-5. The Grade 9 FT items were developed to the new Grade Level Content Expectations (GLCE) for grades 6-8. The test structure for the social studies tests is summarized in Table 2.2.1.7.

**Table 2.2.1.7**  
**Test Structure for Fall 2012 Social Studies Tests**

Grade	# MC Items	# CR Items	# Total Core	# Field Test Item
6	45	0	45	15
9	44	0	44	22

### *Accommodations*

Each operational test is available to students who require accommodations according to their Individual Education Plan (IEP). Tests are available in Braille, large print, and audio CDs. For tests with embedded field test items, Form 1 of the test is the basis for the audio versions. All test forms are converted to large-print, however, Braille tests are typically the Form 1 and may or may not include field test items.

For Fall 2012, unique Braille forms were administered for both Reading and Mathematics. Given the nature of the Reading tests, these tests are not provided as tapes or audio CDs. Tests with accommodations are administered during the same testing window as regular operational tests.

### **2.3. Review of Field Test Items Provided by Development Contractor**

This section provides an overview of the review of field test items provided by the development contractor. Specific item review process at various test development stages are described in other sections of Chapter 2.

#### **2.3.1 Tabulations of Item Characteristics**

Tables 2.2.1.1 to 2.2.1.6 and Tables 2.8.1.1 to 2.8.1.6 provide the tabulations of item characteristics by assessments, including content area, type of item (core, extended core, etc.).

#### **2.3.2. Item Specifications**

MEAP employs *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) as a primary source of guidance in the construction, field testing, and documentation of tests. The introduction to the 1999 *Standards* best describes how those *Standards* were used in the development and evaluation of MEAP tests: Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and acceptability cannot be determined by using a checklist. (*Standards*, p. 4) Thus, the terms ‘target’ and ‘goal’ are used when referring to various psychometric properties of the tests. For example, while it is a goal of test development for each high school test to have a reliability coefficient of .90 or greater, it is not our intention to scrap a test with a reliability coefficient of .89. Instead, the test results would be published, along with the reliability coefficient and associated standard error of measurement.

#### **2.3.3. Item Statistics**

Because the MEAP tests are used in making individual decisions about students, they must be very reliable, particularly at cut points (the score points that separate adjacent achievement categories).. Target reliability coefficients of .90 (or higher) are therefore set for each test. Other psychometric properties include item difficulty, item discrimination, and differential item functioning. General statistical targets are provided below:

##### *For Multiple-Choice (MC) Items*

Percent correct: between 25 and 95 percent  
Point biserial correlation with total score: .20 or greater  
Mantel-Haenszel: Few Category C items

##### *For Constructed-Response (CR) Items*

Difficulty: any level as long as all score points are well represented  
Correlation with total score: .30 or greater by-group statistics and score distributions  
Generalized Mantel-Haenszel: Few chi-square significant at .05 level of alpha

It should be pointed out that the point biserial correlations for MC items and the correlations for CR items assume embedded field testing and employ the base test total score, which is independent of the field tested item. These correlations refer to total scores with the influence of the item in question removed.

#### **2.3.4. Differential Item Functioning**

Items that disadvantage any identifiable subgroup of students are said to be biased and detract from the validity of the tests. While only human judges can determine whether or not an item is biased, item statistics can serve as a tool to help judges in their decisions. After field testing, the BRC reviews item statistics that detect differential item functioning (DIF). Specifically, Mantel-Haenszel statistics are used as measures DIF.

#### **2.3.5. Data Review**

As mentioned previously, once field testing is completed contractor staff analyze results and present them to the CAC and BSC committees for review. The CAC and BSC committees receive training on how to interpret the statistics presented to them. Statistical flags are also set for the committee members to draw attention to items that may not be performing as expected. The DIF statistics comparing Males with Females and White with Black students is converted into a user-friendly flag of A (no DIF), B (possible DIF), or C (probable DIF). Low point-biserial correlations ( $<.25$ ) and  $p$  values ( $<.25$ ) are signaled as well to draw committee members' attention to items that do not connect well with other items on the test or may be too difficult.

Using DIF information, Michigan content standards (CAC committee only), and expert judgment, committee members vote to accept or reject items. For more detailed information concerning DIF statistics, please refer to Chapter 10, Section 10.3.1.

### **2.4. Pre-Field-Test Item Review**

#### **2.4.1. Contractor Review**

The item-writing process begins during the summer. Item writing during the summer of 2012 was done for future MEAP assessments beyond MEAP 2012. Data Recognition Corporation (DRC) conducts either one or two training sessions, depending on the content area, to train teachers to write high quality items. The item writers work on items between those meetings, if applicable, with DRC providing feedback as much as possible. Once item writers have submitted items back to DRC content staff, the items are sent through various internal review rounds

Once this is completed, DRC prepares the items for the first committee review meetings, which typically occur in early fall. The committees consist of a face-to-face Bias/sensitivity Committee (BSC) Review meeting (involving 10-15 Michigan educators) and a Content Advisory Committee (CAC) Review meeting (involving roughly the same number of educators). The BSC meetings may last from one to three days, while the CAC meetings typically last from two to four days. Groups are often broken out by grade span which allows grade-level educators to spend more time focusing on the nuances of each item and tweak them as they see fit.

After this round of reviews, items are approved by BAA staff. DRC then incorporates all changes into items and prepares them for field testing. When all changes have been incorporated into the items, the

items are ready to field test. They are placed on forms, reviewed internally again, sent to the BAA for review, returned to DRC for editing and revision, returned to the BAA for sign-off, and reviewed for overall quality control one final time before they are sent to DRC for printing.

## **2.4.2. BAA Review**

Michigan item writers draft test items in accordance with specifications approved by BAA test development staff. DRC staff then review the items internally and confer with BAA test development staff for additional feedback and approval. BAA test development and DRC staffs work very closely to ensure items are of the highest possible quality. These reviews, both internal and by BAA test development staff, meet the following criteria: skill, content, relevance, accuracy, format, and bias. Detailed criteria reviewed by BAA are provided in section 2.1.4.

## **2.5. Field Testing**

### **2.5.1. Field Testing Design**

BAA conducts field testing by embedding matrix-sampled field-test items within multiple forms of operational assessments such that each field-test item appears on only one operational form, unless otherwise decided by BAA. Field testing is conducted in such a way as to minimize the number of answer documents that must be produced (e.g. different answer documents are required when field testing open-ended items versus multiple choice items).

### **2.5.2. Field Testing Sampling**

Because BAA employs pre-equating (the use of fixed item parameters from field testing) as a critical part of its equating and scaling methodology in grades 3-8, it is critical that field test items be calibrated with operational items in such a way that the pre-equated item parameters represent those parameters that would result were the field test items administered to all students. To assure this outcome, the multiple operational forms were randomly distributed to buildings using a stratified random sampling plan. The sampling plan identifies three strata within which a random assignment of Forms 1 through 5 should be made. Inherent in the design is that every building in the State will be identified as a Form 1, 2, 3, 4, or 5 building. No building should have to deal with any Initial Form that is different from this assignment (Makeup and Accommodated forms, will not carry the building's designated Form number).

Stratum I includes Detroit, Utica, Grand Rapids, and the Education Achievement Authority (EAA). Detroit has 27 buildings with ninth-grade students, who would be expected to take the social studies assessment. Grand Rapids has six, and Utica has seven. This is a total of 40 schools in these three districts that house ninth-grade students, or eight schools per form of the social studies tests. EAA took over 12 of Detroit schools so that we officially add EAA to Stratum I. Because of the diverse characteristics of these district populations, we select five buildings in the Stratum to receive each given form of the ninth-grade social studies test booklets. There are so many elementary and middle schools in these districts that a random assignment of forms should ensure that a representative cross section of students will take each form.

Stratum II consists of seven districts with demographics that warrant special consideration for receiving more than one form. Flint, Lansing, Plymouth-Canton, Saginaw, Highland Park, Muskegon Height, and Holland are each notable for their unique demographics. They should receive specific consideration for

forms distribution, given multiple forms (randomly assigned) for each district but again having only one form assignment for a building. Dearborn schools are also in this stratum but will require slightly different handling due to the high proportion of Arabic students who are also ELLs in selected schools. After consulting internally and with Shereen Tabrizi of Dearborn Public Schools, it was determined that the following schools within the Dearborn district will be pre-designated as form 1 schools: Maples Elementary, McDonald Elementary, Salina Elementary, Salina Intermediate, Lowrey Elementary, Lowrey Middle School, Iris Becker Elementary, McCollough Elementary. The rest of the Dearborn schools are to be assigned forms 2-5. In addition, 100 extra form 1s, per grade and content area, were sent to the Dearborn Schools Central Office care of Shereen Tabrizi. Dearborn was instructed to be proactive in the ordering of the other materials they will need for the ELL accommodations from MI.

Stratum III consists of all of the remaining districts. Each district should receive only one form of each test across all buildings within a given district. In other words, a district would receive the same form number for all of the assessments. These numbers were assigned at random. For example, Ann Arbor Schools might be a "Form 3" district, with Form 3 being assigned to all grades of ELA, mathematics, science and social studies across the entire district.

It is critical that within strata 1 and 2 that all sampling be done without replacement according to the defined parameters of the strata. That is, the sampling plan will be carried out in such a way that the forms are distributed according to stratum specific rules while also ensuring that all forms are distributed before a second school within a district can receive a given form that had already been assigned to another school within that district.

Stratum II was built upon the following variables where the MEAP was administered in previous years (some schools and grades did not previously administer the MEAP):

- Percent limited English proficiency in the school building
- Percent ethnicity in the school building
- Total number of students

The stratified random form assignment process is summarized as follows. For each grade, for each subject (take Mathematics for example):

1. Compute "population values" across all forms for all variables of interest: percentage of limited English proficiency; percentages of white, black, Hispanic and others for ethnicity.
2. Randomly assign form to each school.
3. Compute the same set of statistics (see step 1) for each form; also compute the sample size for each form
4. Compute the sum of the absolute difference between population values and sample values.
5. Repeat steps 2 through 4 for 5,000 times and pick out one sample - criteria are: (1) smallest difference between population values from step 1 and sample values from step 3 across the variables; (2) similar sample sizes across forms

The results of the matrix sampling are displayed in Appendix A. Statistics in Appendix A show that the sample sizes across forms are relatively even, and “form values” are very closely matched with “population values”. The stratified random form assignment has been successfully achieved.

## **2.6. Data Review**

After field-test administration, item analyses were conducted to prepare data for two more rounds of reviews: bias/sensitivity review and content review. This section describes data for the two reviews and these two post-field-test reviews.

### **2.6.1. Data**

All field-test items were embedded in the live test forms for each test. After the calibration of live test forms, field-test items were calibrated and put onto the same scale as the live operational items. Appendix B lists all the statistics created for field-tested items. The statistics for each field-test item can be summarized into nine categories.

1. General test information: test name, subject, grade, level;
2. Administration related information: year cycle, administration year, released position;
3. Specific item information: MEAP item ID, CID, item type, answer key, maximal score, maturity, item function, character code, number of forms the item appears on, form numbers, test position, n-count (total, male, female, white, and black students), percent for each comment code, percent for each condition code;
4. Content-related information: strand, benchmark, grade level expectation, depth of knowledge, domain, scenario;
5. Option analysis: percent for each option and each score point (total, male, female, white, and black students), p-value or item mean (total, male, female, white, and black students), adjusted p-value, difficulty flag, item standard deviation, item-total correlation, biserial/polyserial correlation, corrected point-serial correlation, item-total correlation flag, option point-biserial correlation, flag for potential miskeying;
6. DIF analysis: Mantel Chi-square, Mantel-Haenszel Delta and its standard error, signed and unsigned SMD, SMD signed effect size, DIF category, and favored group for male vs. female comparison and white vs. black comparison (Computation of DIF statistics is described in detail in Appendix C);
7. IRT parameters: b-parameter and its SE, step parameters and their respective SE, item information at cut points;
8. Fit statistics: mean-square infit, mean-square outfit, mean-square fit flag, misfit level;
9. Data for creating plots: conditional item mean for decile 1 to 10 for each student group (total, male, female, white, and black students) for creating conditional mean plots, 5<sup>th</sup>, 25<sup>th</sup>, 50<sup>th</sup>, 75<sup>th</sup>, 95<sup>th</sup> percentile for creating Box & Whisker plot for each student group (total, male, female, white, and black students) for each option and each score point.

These statistics were created by BAA psychometricians and sent to DRC for creating item labels for bias/sensitivity review and content review.

## 2.6.2. Statistics Prepared for Review Committees

Statistics from item analyses of field-test items were used to create item labels for the Data Review meetings. Different sets of statistics were prepared for MC and CR items for committee review. Table 2.6.2.1 displays all the statistics prepared for MC items for committee review. These include six categories.

1. General administration information: test name, grade, subject, and administration time;
2. Item general information: CID, maturity, forms and positions;
3. Item specific information: item type, key, p-value, n-count, Rasch difficulty, difficulty flag, point-biserial correlation, point-biserial correlation flag, fit flag, option quality flag;
4. Breakout group descriptives and optional analysis: percent of students selecting each option and omit, option point-biserial correlations, and n-count for all and subgroups: male, female, white, and black students;
5. Differential Item Functioning: flag, and favored group for male vs. female and white vs. black;
6. Review decision;

Table 2.6.2.2 displays all the statistics prepared for CR items for committee review. These include seven categories.

1. General administration information: test name, grade, subject, and administration time;
2. Item general information: CID, maturity, forms and positions;
3. Item specific information: item type, maximal score point, adjusted p-value, item mean, n-count, Rasch difficulty, difficulty flag, item-total correlation, item-total correlation flag, fit flag, score point distribution flag;
4. Breakout group descriptives and score point distribution: percent of students obtaining each score point and omit and n-count for all and subgroups: male, female, white, and black students, omit point-biserial correlation;
5. Invalid code distributions: total invalid scores, frequency of students at each invalid code;
6. Differential Item Functioning: flag, and favored group for male vs. female and white vs. black;
7. Review decision;

All statistics prepared for committee review of MC and CR items are explained in Appendix D. When the p-value for a MC item, or adjusted p-value for a CR item, or Rasch difficulty was out of the desired range, a difficulty flag was shown. When point-biserial correlation for a MC item or item-total correlation for a CR item was out of range, a point-biserial or item-total correlation flag was shown. If the mean square infit or outfit was out of desired range, a mean-square infit or outfit misfit flag was produced. If the DIF level for male vs. female or white vs. black comparison was higher than moderate, a DIF flag was turned on. When options did not function well or score point distribution was abnormal, a miskey flag was produced. The criteria used for flagging an MC or CR item are presented in Table 2.6.2.3.

**Table 2.6.2.1**  
**Item Label for a MC Item**

**MEAP    Grade: 3            Subject: Math            Admin: 2012**

**CID:** 100000160320

**GLCE:** M.PS.02.02

**Form:** 0903

**Position:** 9

- |                                               |
|-----------------------------------------------|
| <input type="checkbox"/> Accept as is         |
| <input type="checkbox"/> Reject               |
| <input type="checkbox"/> Accept with revision |

**Passage:**

**Item Information**

<b>Type:</b> MC	<b>P-value:</b> 0.82	<b>Difficulty Flag:</b>	
<b>Key:</b> C	<b>N-count:</b> 15523	<b>PB Correlation:</b> 0.32	<b>PB Correlation Flag:</b>
	<b>Maturity:</b> FT	<b>Option Quality Flag:</b>	

**Breakout Group Descriptives and Option Analysis**

		N-count	Percent of Students Selected Option				
			A	B	C	D	Omit
<b>Group</b>	<b>All</b>	15523	11	7	82 *		0
	<b>Male</b>	7833	12	6	82		0
	<b>Female</b>	7690	10	8	82		0
	<b>White</b>	10202	8	4	87		0
	<b>Black</b>	3725	17	14	69		0
<b>Option PB Correlations</b>			-0.17	-0.28	0.32		-0.04

**Differential Item Functioning**

<b>Reference/ Focal Group</b>	<b>Male/ Female</b>	<b>White/ Black</b>
<b>Flag</b>	A	B
<b>Favored Group</b>		WHITE

<b>GLCE Description:</b>	Compare, add, subtract lengths
------------------------------	--------------------------------

*\*Note: IRT statistics were not provided during the most recent data review (February 2012).*

**Table 2.6.2.2**  
**Item Label for a CR Item**

**MEAP      Grade: 3      Subject: Reading      Admin: 2012**

**ID:**            **100000234324**                      **Maturity:** FT  
**Form:** 0905  
**Position:** 40  
**Passage:** Grandfather Buffalo

- ☐ Accept as is  
☐ Reject  
☐ Accept with revision

**Item Information**

<b>Type:</b> CR	<b>Adj. P value:</b> 0.58	<b>Difficulty Flag:</b>
<b>Max:</b> 3	<b>Item Mean:</b> 1.74	<b>Item-Total Corr:</b> 0.47 <b>Item-Total Corr Flag:</b>
	<b>N-count:</b> 1981	<b>Score Point Dist. Flag:</b>

**Breakout Group Descriptives and Score Point Distributions**

		N-count	Item Mean	Percent of Students at Each Score Point							
				0	1	2	3	4	5	6	Omit
Group	All	1981	1.74	13	27	30	29				0
	Male	1006	1.66	15	28	31	26				0
	Female	975	1.82	11	26	29	33				0
	White	1464	1.85	10	27	30	33				0
	Black	308	1.33	23	30	29	15				1
Omit PB Correlation				-0.08							

**Condition Code Distribution**

Percent of Students at Each Condition Code			
A	B	C	D
0	0	0	13

**Differential Item Functioning**

Reference/ Focal Group	Male/ Female	White/ Black
<b>Flag</b>	AA	BB
<b>Favored Group</b>		WHITE

<b>GLCE Description:</b>	Retell in sequence the major idea(s) and relevant details of grade-level narrative and informational text.
--------------------------	------------------------------------------------------------------------------------------------------------

*\*Note: IRT statistics were not provided during the most recent data review (February 2012).*

**Table 2.6.2.3  
Flagging Criteria**

Statistic	Flag	Flag Definition	Flag Field
PVAL PVAL ADJPVAL	PL PH BL BH	For MC 4 options, if p-value LT .3 (PL) or GT .9 (PH) For MC 3 options, if p-value LT .38 (PL) or GT .9 (PH) For CR items, if adj. p-value LT .10 (PL) or GT .9 (PH)	DIFFICFL
ITOT	CL	If item-total correlation LT 0.25 (CL)	ITOTFL
DIF_MF DIF_WB	A B C  AA BB CC	For MC items: A: If either  MH Delta  is not significantly GT 0 ( $p < 0.05$ , using either MH-Chi-Sq or standard error of MH Delta) or if the  MH Delta  is LT 1 B: If  MH Delta  is significantly GT 0 and is either GE 1 and LE 1.5 or is GE 1 but not significantly GT 1 ( $p < 0.05$ , using standard error of MH Delta) C: If  MH Delta  is both GT 1.5 and significantly GT 1 ( $p < 0.05$ , using standard error of MH Delta) For CR items: AA: If the Mantel Chi-Sq is not significant ( $p > 0.05$ ) or the  Effect Size  (ES) of SMD LE 0.17 BB: If the Mantel Chi-Sq is significant ( $p < 0.05$ ) and the  ES  is GT 0.17 but LE 0.25 CC: If the Mantel Chi-Sq is significant ( $p < 0.05$ ) and the  ES  is GT 0.25	DIF_MF DIF_WB  Categories A and AA are not displayed in flag field
A, B, C, D M, S5, S6, O  APB BPB CPB DPB OPB	H L P (O) O N B	For MC items: If the keyed option is not the highest percentage (H) If any option LE 2% (L) If any non-keyed option pb-corr GT 0 (P), or if omit pb-corr GT 0.03 (O) If the keyed option pb-corr LT 0 (N) For CR items: For CR, if omit pb-corr GT 0.03 (O) For CR, if any score point LT 0.5% (L) For CR, if omit GT 20% (B)	MISKFL

*\*Note: IRT statistics were not provided during the most recent data review (February 2012).*

**Meaning of Flags:**

PL ... p-value low	H ... highest percentage is not a keyed option
PH ... p-value high	L ... low percentage of any option
CL ... correlation low between item and total	P ... positive pb-correlation for any non-keyed option
A or AA ... no or negligible DIF	N ... negative pb-correlation for the keyed option
B or BB ... moderate DIF	O ... omit has a positive pb-correlation
C or CC ... substantial DIF	B ... blanks are over 20%

### **2.6.3. Data Reviews**

#### **2.6.3.1. Bias/Sensitivity and Content Advisory Committee Review**

DRC planned and conducted Bias/Sensitivity committees (BSC) reviews followed by Content Advisory committees (CAC) reviews on field tested items that were flagged either because of Differential Item Functioning (DIF) or any other content related item property (see flagging criteria on Table 2.6.2.3). The goal of these committees was to identify items that are eligible to be used as scorable items in future operational assessments. In during these reviews, items may be either: a) accepted “as is” or b) rejected,

The BAA psychometricians scored the field test items and provided data analyses that are necessary for the BSC and CAC reviews. DRC assembled all materials for the meetings including the items, data and analyses of the items, agenda, training materials, security agreements, sign-in sheets, and the necessary records for committee sign-off on each item. DRC conducted the meetings after obtaining approval from the BAA on the agenda, training, and process. The first part of the item book contained items that were flagged for BSC reviews and the second part contained items flagged for CAC reviews. Some of the items for BSC reviews may have been also flagged for CAC reviews. DRC completed a comprehensive report summarizing the results of these meetings including attendance, decisions on each item, data and statistical analyses, and final disposition of the items.

DRC prepared items following field testing for reviews by a Bias/Sensitivity Committee and a Content Advisory Committee. Each committee met in face-to-face meetings lasting anywhere from a half day to one or more days depending on the number of flagged items. The reviews were guided by checklists to ensure that the items met the criteria for inclusion in the item bank and for potential use on future examinations. DRC reviewed the flagging and reviewed criteria with the BAA to be sure that all nuances of acceptability were captured correctly. The item statistics for each item were presented, along with a general orientation of interpretation and use of the data in item approval.

#### **2.6.4. Item Revision Procedures**

It is DRC’s policy to leave post-field test items as intact as possible since they have data attached. Making major changes can negate that data. However, there are circumstances where the item may be revised.

Generally, the data review committee participants examine the items and either accept them as is or reject them. Occasionally, committee members suggest minor revisions that could improve the clarity or quality of the item. BAA test development must approve of any changes to the item, and if the committee or the BAA believes that significant changes are required to improve the item, it is rejected as ready for operational use.

The committee's recommendations are brought back to DRC and entered into the system. At this time, the field tested items are available for use on operational forms. Items selected for operational use will be composed and reviewed by DRC staff, then sent to the BAA test development staff for review. Minor changes may be requested as the items are considered as a group. Once DRC has made these revisions and the BAA has approved the form, the DRC quality assurance team reviews the form one final time before it is sent to the printer.

## **2.7. Item Banking Procedure**

The Michigan Item Banking System is a secure web-based application that dynamically supports in one system:

- 1) all item development processes throughout the entire life cycle of an item from assignment through retirement;
- 2) all test development processes from blueprint design through test map generation and approval, and subsequent uploading of item statistics for a test administration;
- 3) item maturity and version control throughout the item development and maturation cycles, controlling item availability within specific item pools (pilot testing, field testing, or operational)
- 4) all state-level summative assessment programs.

**Access** - Access to the Item Banking System (IBS) is controlled by Tivoli Single Sign On authentication. Access to items within the IBS is based on user role permissions, item maturity, and specified assessment program (e.g. MME), content area (e.g. Science) and grade level permissions.

**Item Assignment** - The item development process begins in the IBS with the assignment of an item to a specific item writer. Item assignments are based on item inventory and blueprint design. The item assignment specifications include the content expectation being measured, item type, taxonomy level (DOK for MME items), and due date. When the item assignment is submitted, the system assigns a unique Item ID to the item. The item will retain this unique ID through its life cycle. The maturity of the item will be updated as it progresses through its life cycle, and the version of the item will be updated with each change to the item. Each version of the item is retained and viewable within the IBS history. The item writer can only access items assigned to them, and only in the item's submitted state.

**Item Development** – When the Item Writer submits the item, the Content Lead can accept the item, request further revision by the Item Writer, or reject the item. Once the

Content Lead accepts the submitted item, they can make further revisions to the item text. If there is a graphic request the item will be routed to the Composition Team to create the item graphic(s). When the graphic requests are fulfilled, the item will be routed back to the Content Lead to review the item and graphics. If revisions to the graphic are needed, this graphic revision and review process will continue until the Content Lead accepts the item for Committee Review.

Committee Members work within the Item Banking System to preview each item and provide their feedback with a recommendation for acceptance, revision, or rejection. Committee Members are only able to view the items assigned to their committee, and only their own feedback. A Committee Facilitator is able to review all committee member feedback and initiate discussion on any item where there is not agreement. The Committee Facilitator will enter the consensus comment into the IBS. All committee feedback is also retained and viewable in IBS.

Following the Initial Committee reviews for BSC and CAC, the Content Lead can accept the item as is; reject the item, flagging as Do Not Use (DNU), which will prevent it from progressing through the system; or edit the item based on committee feedback, and route for graphic revisions if needed. Once the Content Lead has accepted the item, it is routed to a Composition Editor who reviews the item in the IBS for proofing and ensuring that the meets the Style Guide specifications. The Editor can approve the item or suggest revision, but cannot alter the item. The Content Lead will determine whether to make the suggested revisions, but the Editor feedback is retained in the IBS.

**Item Banking** - Once the Content Lead and the Editor have approved the item content (there is no layout at this point), the Content Lead “banks” the item by routing it to the appropriate item pool, updating the maturity as Ready for Pilot Test or Ready for Field Test. The item never leaves the IBS. This process maintains maturity and version control of the item, while removing the item editing and revision process from the critical path of test development and form production.

The item is available for use in the applicable inventory pool based on its maturity. Item statistical data upload into the IBS will advance the maturity of the item to Pilot Tested or Field Tested and route the item for Data Review. The data review process is similar to the Initial Item Review process defined above with the addition of the statistical data being available in the item bank for committee member review. An item may be routed to the Operational pool by the Content Lead following Data Review; the item may undergo suggested revision and be routed for Re-Field Testing; or the item may be rejected and flagged as DNU which removes it from any item pool availability.

**Test Development** – The IBS provides the functionality for a Content Lead to build a test blueprint inside the item bank, specifying the number of forms, quantity and type of

items by content expectation, item function (common, matrix, or field test), and identify equating or linking items for the test map.

The IBS will generate a test layout showing the content expectation, item type, and item function in each test position. The Content Lead can rearrange the item positions based on the preceding criteria. Once the Content Lead approves the test layout, the system will select the items to fill the test map based on the blueprint criteria and the selection algorithm.

The test map is then available for review and approval by the Content Lead. The Content Lead can rearrange or replace items during their review process. The most recent item statistics based on the administration type (standard, accommodated, make-up) will be displayed in the test map, and the system will generate for each form:

- a statistical summary for each test form including summary statistics for the adjusted p-value, item-total correlation, the three parameters and their standard errors, if available;
- Summary Frequency for Scoring keys, DIF ranges, and item types;
- Item Statistics Detail including Adjusted p-value, Item-Total Correlation, each of the three parameters and their standard error, if available;
- Test Characteristic Curve, Test Information Curve, and Test Standard Error Curve compared to the Base Curves selected by the Psychometrician.

**Psychometric Approval of Test Map** – Once the Content Lead has approved the items in the test map; the Psychometrician will receive notification from the IBS that there is a test map pending their review and approval. The Psychometrician can approve the test map as is, or recommend revision. The Content Lead or Psychometrician can search the item bank to identify items that match the criteria to improve the test map. The Content Lead can replace items in the test map until both the Content Lead and Psychometrician have approved the test map. At that time the test map is “locked down” and no additional changes may be made to the test map.

**Creation of Test Forms** – Once the test map has received both the Content Lead and Psychometric approval, the system will generate an export of the item elements (stem, options, and graphics) for each unique item in the test map.

The Composition Team will create a OnePer for each item. The OnePer is a single page layout presentation for each item (one item per page) to represent how that item will be displayed each time it appears in a test form. Each unique item in the test map will only be formatted once. This ensures consistency of item presentation across forms and test cycles.

The OnePers for the test map are uploaded into the IBS. The Content Lead will then compare the OnePer against the IBS to ensure content accuracy. Following Content Lead approval, the Editor will then review the OnePers to ensure item integrity with the Item Bank content. When the IBS system receives both OnePer approvals for the test map, the system will export the pre-composition materials (approved OnePers) in sequence for each form in the test map, for creation of the printed test booklets from the sequenced OnePers.

At this point in the process, the review of the individual test forms becomes external to the IBS. The cycle resumes in the IBS with the upload of Item Statistics after test administration, continuing the cycle of data review, items advancing in maturity, and being selected for appearance in a test map at the appropriate maturity level.

## **2.8. Construction of Operational Test Forms**

The Michigan Bureau of Assessment and Accountability (BAA), Measurement, Inc., and DRC work collaboratively to develop and construct the operational test forms used to support the MEAP program. In 2012, operational test forms were developed for the following subject areas:

- Reading: grades 3-8
- Writing: grades 4 and 7
- Mathematics: grades 3-8
- Science: grades 5 and 8
- Social Studies: grades 6 and 9

Test form development entails the following steps:

- Review the assessment blueprints for the operational assessments
- Select assessment items to meet the content and process specifications of the assessment blueprints
- Assess the statistical characteristics of the selected assessment items
- Review and approve test forms

The following sections discuss essential aspects, include guidelines, and identify important references to follow through the four-step process.

### **2.8.1. Design of Test Forms**

The following section describes the detail of test form design for the MEAP program.

#### **2.8.1.1. Review the Assessment Blueprints for the Operational Assessments**

As the name implies, the assessment blueprints identify the content and types of items to be included on the operational forms. These specifications include benchmark and

content targets (limits), general indicators of difficulty and other psychometric characteristics, as well as general physical indicators such as passage length and artwork parameters.

All MEAP assessments are designed to assess higher order thinking skills. Most items in all subject areas focus more on comprehension and application than on simple recall or recognition. Indeed, specifications for each assessment clearly include admonitions to avoid simple recall of trivial or unrelated facts.

The MEAP assessments use three different types of items; multiple choice (MC), constructed response (CR- Reading and Writing only), and writing prompts (Writing only). Each item is aligned to a specific domain, standard, and objective. The alignment information is used during the forms construction process to help ensure the forms meet the blueprints.

In addition to the operational items, each form also includes embedded field test items. To minimize location effects, field test slots occur in multiple locations in mathematics, science, and social studies. Due to the nature of the ELA materials, a separate section of the assessment is reserved for the field test slots. From year to year, the number of field test slots on a form typically remains constant.

The following sections outline the assessment blueprints for all subject areas and grade that were used to construct the fall 2012 test forms.

### ***Mathematics Content for Grades 3-8***

Mathematics items for grades 3-8 are identified as core or extended core. Core refers to the content most commonly taught at a grade level. Extended core refers to content that is typically taught at grade level but is narrower in scope and/or is supportive to the core. Since the grade 3-8 assessments are administered in the fall, each assessment is based on the grade level content expectations (GLCEs) from the previous grade. Table 2.8.1.1 shows the breakdown of the grade 3-8 mathematics assessments.

**Table 2.8.1.1**  
**Fall 2012 Mathematics Grade 3-8 Assessments**

Grade	# Core			# Total Core	# FT		# Field Test Per Form
	MC	CR	Extended		MC	CR	
3	32	0	13	53	8	0	8
4	40	0	13	59	9	0	9
5	32	0	20	54	10	0	10
6	38	0	14	60	9	0	9
7	38	0	15	62	10	0	10
8	40	0	11	49	8	0	8

***English Language Arts Content for Grades 3-8***

The Reading assessment consists of multiple parts. Each assessment consists of 3-4 reading texts of varying lengths. At each grade, the assessments have one paired text along with one independent text. At grades 4-8, there is an extra linking passage to allow for measuring student growth. Finally, each form contains embedded field test items.

For 2012, the Writing test form assessed writing via one writing prompt and a series of multiple choice questions. Tables 2.8.1.2 and 2.8.1.3 summarize the fall 2012 Grade 3-8 ELA assessments.

**Table 2.8.1.2**  
**Fall 2012 Reading Assessments Grade 3-8**

Grade	Reading Core		Reading Field Test		Total
	MC	CR	MC	CR	
3-8	30	1	16	2	51

**Table 2.8.1.3**  
**Fall 2012 Writing Assessments Grades 4 and 7**

Grade	Writing Core		Writing Field Test		Total
	MC	CR	MC	CR	
4, 7	NA	NA	8	1	9

***Science Content for Grades 5 and 8***

In science, items are classified according to strand. In addition to the three major science subject areas (Life, Earth, and Physical science), each form must also include a set number of items that address the process skills of Constructing and Reflecting. Tables 2.8.1.4 and 2.8.1.5 summarize the science assessments for grades 5 and 8, respectively.

**Table 2.8.1.4**  
**Grade 5 Science Assessment**

Discipline	OP MC Items	FT MC items	Total Items	Total Points
Science Processes	13	3	16	13
Life Science	7	3	10	7
Earth Science	12	3	15	12
Physical Science	16	3	19	16
Total	48	12	60	48

**Table 2.8.1.5**  
**Grade 8 Science Assessment**

Discipline	OP MC Items	FT MC items	Total Items	Total Points
Science Processes	13	3	16	13
Life Science	13	3	16	13
Earth Science	14	3	17	14
Physical Science	13	3	16	13
Total	53	12	65	53

***Social Studies Content for Grades 6 and 9***

Table 2.8.1.6 and 2.8.1.7 summarize the blueprint for the social studies assessments

**Table 2.8.1.6**  
**Social Studies Blueprint: Grades 6**

Strand	MC Items	Total Items	Field Test Items	Total Points
History	19	19	*	19
Geography	7	7		7
Civics	10	10		10
Economics	7	7		7
Knowledge, Processes, Skills	2	2	*	2
Total	45	45	15	45

*\*Each field-test form consists of items from the appropriate strands as determined by DRC and BAA.*

**Table 2.8.1.7**  
**Social Studies Blueprint: Grades 9**

Strand	MC Items	Total Items	Field Test Items	Total Points
History	23	23	*	23
Geography	13	13	*	13
Civics	3	3	*	3
Economics	5	5	*	5
Knowledge, Processes, Skills	0	0	*	0
Total	44	44	15	44

*\*Each field-test form consists of items from the appropriate strands as determined by DRC and BAA.*

**2.8.2. Item Selection**

In addition to content coverage requirements, the forms must also meet certain statistical targets. These targets are outlined in the next three sections.

### **2.8.2.1. Select Assessment Items to Meet the Assessment Blueprints**

Following field testing, the items are submitted for review to both the Bias Review Committees (BSCs) and the Content Advisory Committees (CACs). These committees, composed of Michigan educators and Michigan citizens, sort the field tested items and identify which items are eligible for inclusion in the operational item pool. There is a separate pool for each subject area and grade level assessed. It is from these pools that items are selected to meet the requirements outlined in the assessment blueprints.

Test forms are developed using approved MEAP items. In addition to overarching content requirements for each test form developed, content experts and psychometricians consider requirements related to subdomains, graphics and other visual representations, passage and content dependent items, and clueing concerns.

### **2.8.2.2. Assess the Statistical Characteristics of the Selected Assessment Items**

The statistical process begins with the work of the Data Review Committees, both the BSC and the CAC post field test. The committees evaluate the field test items using item statistics from classical measurement theory and item response theory models. From the work of these committees, a pool of items that are eligible to be used in constructing the operational forms is identified.

Because the MEAP assessments are used in making individual decisions about students, they must be very reliable, particularly at cut points (the score points that separate adjacent achievement categories). The targeted reliability coefficient is .90 (or higher) for each assessment.

Other psychometric properties include item difficulty, item discrimination, and differential item functioning. General item and form level statistical targets are provided below:

#### *For Multiple-Choice (MC) Items*

- Percent correct:  $.25 \leq p\text{-value} \leq .95$
- Point biserial  $\geq .25$
- Mantel-Haenszel: Few Category C items<sup>1</sup>

#### *For Constructed-Response (CR) Items*

- Difficulty: any level as long as all score points are well represented
- Correlation with total score: .30 or greater by group statistics and score distributions
- Generalized Mantel-Haenszel: Few chi-square significant at .05 level of alpha

---

<sup>1</sup>For category C items, D's absolute value is significantly greater than or equal to 1.5

To help ensure adequate coverage of a full range of achievement on the operational assessments, the draft forms are evaluated to see whether the following targets are met. As necessary, items are replaced on the draft forms until this distribution is approached.

<b>Rash Item Difficulty</b>	<b>% of items</b>
-2.00 to -1.00	25
-0.99 to 0.00	25
0.01 to 1.00	25
1.01 to 2.00	25

Even with careful test form development, it is usually not possible to create alternate forms that are exactly equal with respect to difficulty. The MEAP assessments are being analyzed using the Rasch Partial Credit Model.

As the MEAP test forms are assembled, spreadsheets are used to track the statistics and other metadata (e.g., alignment) for the selected assessment items. Both classical and Rasch statistics are included. The statistics listed on the spreadsheets include item *p*-values, correlations, and Rasch item difficulties for multiple-choice items and item means, standard deviations, correlations, and Rasch step difficulty estimates for constructed-response items.

The above two steps require an iterative process to create test forms that are a combination of the content and statistical information. Working together, DRC psychometricians and content experts replace items until both groups are satisfied with the forms. Through this iterative process of item selection, item content takes precedence over statistical characteristics.

### **2.8.2.3. Review and Approve Test Forms**

BAA test development staff create test forms and test maps for each assessment. During the test design process, staff include open slots for embedded field test items. The BAA test development and psychometric teams review test forms to determine whether both content and statistical requirements are met.

Guidelines for test forms review include:

- Confirm that all assessment items were accepted by the BAA and the committees
- Confirm that all blueprint requirements are met
- Confirm that all content considerations including content/skill/topic balance, correct keys, no clueing, and correct graphics are met.
- Confirm that the item and mean difficulty levels are accurate and meet requirements
- Confirm that the assessments cover a full range of achievement levels

As necessary, BAA will replace items that are identified by BAA as problematic, either from a content or psychometric perspective. As items are replaced, the match of the newly revised test form to the specifications is updated and reviewed. This process continues until BAA has approved each form.

## **2.9. Accommodated Test Forms**

A testing accommodation is a change to the testing environment to assist a student with special needs so that assessments mirror instruction as much as possible without invalidating test results. District and school testing coordinators are responsible for communicating information about testing accommodations to test administrators and other interested individuals. Information about testing accommodations is also included in the test administrator manuals.

The decision to use a particular accommodation with a student should be made on an individual basis and should take into consideration the needs of the student and whether the student routinely receives the accommodation in classroom instruction and testing. If a student receives special education services, all accommodations must be documented in the student's individualized education program (IEP).

Typically, accommodations allow for a change in one or more of the following areas:

- Presentation format
- Test setting
- Scheduling or timing
- Response format

### **2.9.1. Special Order Accommodated Testing Materials**

The following accommodated testing materials are provided for MEAP: Braille, Enlarged Print, Oral Administration (except Reading) and Audio Translations.

### **2.9.2. Braille**

All test items are screened for adaptability to Braille. If an item not suitable for Braille is selected for use on a base-test form, an appropriate item would be substituted on the Braille form or the item would be dropped from the Braille form.

### **2.9.3. Enlarged Print**

An enlarged print version of Form 1 is created as a standard accommodation for students who need the accommodation as defined in their IEP or Section 504 Plans. Students who use this accommodation have their answers transferred onto a regular answer document.

#### **2.9.4. Oral Administration for Mathematics, Science, Social Studies and Writing**

Reader Scripts are created of Form 1s (other than Reading) including the portion of each item that may be read aloud without compromising the construct of the test. For example, if a problem requires students to identify the largest number, the answers would not be read aloud. Reader Scripts must be read by the test administrator exactly as written. Students may also use an audio CD of the scripted test. DVDs are also available and each item is read aloud as it is presented on the screen.

#### **2.9.5. Bilingual Tests**

Translations of the test are made for Spanish and Arabic which are the top language groups represented in the state after English. Students may have tests interpreted on the day of testing for languages where a printed bilingual version is not available.

## **CHAPTER 3:**

### **OVERVIEW OF TEST ADMINISTRATION**

#### **3.2. Test Administration**

For each administration, Measurement Incorporated (MI) designs forms to assist the Assessment Coordinators with the successful receipt and return of test materials. These forms provide security and accountability during the fulfillment and distribution, test administration, and collection process.

Three months prior to the assessment window, MI receives a file from the Bureau of Assessment and Accountability (BAA) containing a list of Assessment Coordinators for MEAP and their shipping and mailing addresses. The BAA also provides a pre-ID file, which contains summary information about the number of students enrolled. These data files contain addresses, enrollments, secure document ID numbers, and other relevant information used to create shipping labels, packing lists, security forms, and control rosters. Below are descriptions of the materials and forms provided in the shipments.

*Picking/packing lists.* MI generates picking/packing lists for each district and school order that contain information specific to each district and school shipment. Materials are packaged by school and sent to the school or district, as indicated by the Assessment Coordinator.

MI provides secure packaging and distribution of materials for the MEAP program to ensure prompt, accurate, and secure delivery of test materials to district Assessment Coordinators. The materials for the MEAP program are packaged by building and then distributed to either the district or the school, depending on the option chosen by that particular district. Secure materials are numbered with a unique barcode identifier to assure 100% accountability during the process of picking, packing, distribution, and return of secure materials.

Quantities of materials provided for each school are calculated based on enrollment counts and overages using business rules that were developed with BAA and documented in the Requirement Specifications for Enrollment Collection and Additional Orders. These specifications include rules about how enrollment counts are generated, overages are calculated, and bill of materials are determined for various items.

MI prints packing lists that indicate the number of test booklets being packaged and the number of shrink-wrapped packages, as well as all non-secure materials being shipped to the district or school. The packing list serves as a list to pick and pack assessment materials associated with each order. After the secure materials are picked, the secure materials barcodes are scanned and a secure materials list is generated which is included with the packing list. Packing lists also serve as inventory lists by providing a place for District and School Assessment Coordinators to check off materials they have received.

*Shipping labels.* MI enters the delivery information for the receiving district or school into the FedEx system and generates FedEx barcode shipping and tracking labels.

*Materials return kits.* MI assembles return kits for district Assessment Coordinators to use in returning materials to MI. The kits contain instructions for preparing completed materials for shipment to MI and all of the necessary control forms, materials identification labels, and pre-assigned FedEx shipping labels and/or bills of lading. Return FedEx tracking numbers are pre-assigned to districts or schools in an effort to facilitate log in of materials. Materials identification labels are color-coded to distinguish boxes containing completed answer documents being returned for scoring from those containing secure test materials (i.e., test booklets). The total number of pre-ID students by grade is reported in Table 3.2.1; the total number of students tested is reported in Table 3.2.2.

**Table 3.2.1  
Number of Pre-ID Students in Fall 2012**

	<b>Mathematics</b>	<b>Reading</b>	<b>Reading Day 2</b>	<b>Science</b>	<b>Social Studies</b>	<b>Writing</b>
3	123314	123011	123011			
4	121037	120673				120498
5	122119	121805		123882		
6	128792	128535			114984	
7	127027	126975				127125
8	126702	126619		128164		
9					118687	

**Table 3.2.2  
Number of Students Tested in Fall 2012**

	<b>Mathematics</b>	<b>Reading</b>	<b>Reading Day 2</b>	<b>Science</b>	<b>Social Studies</b>	<b>Writing</b>
3	109751	108963	108853			
4	108168	107270				107379
5	108084	107443		110777		
6	111622	111114			114667	
7	113950	113806				113850
8	113396	113267		115762		
9					123339	

### **3.3 Materials Return.**

*Shipping.* In order to retrieve materials immediately after testing, while providing maximum flexibility to schools and districts, MI uses FedEx Express 2-Day service for the return of all assessment materials.

The upgrade in service level provides a rapid and consistent flow of material in an effort to meet the stringent time constraints required by BAA for these assessments. This service guarantees delivery of materials no more than two days after pick-up in Michigan, and there is greater probability that all boxes from a district are delivered on the same date.

*Return kits.* Districts are provided with “Return Kits” containing all of the necessary labels and documentation that are used for returning their materials. The tracking numbers of the return labels provided to each district are entered into our internal tracking system database at the time of “Return Kit” production. This process offers an accurate and expedient method of logging materials in upon return to MI.

Materials are prepared for return by the MEAP test coordinator. The directions for packaging and returning test materials are explained in detail in the MEAP Test Administrator Manual supplied to each test coordinator. The coordinator packages the materials and applies the self-adhesive return label that is supplied in the “Return Kit” from their original shipment. The coordinator then calls a toll-free telephone number to arrange for pick-up of their materials.

*Pick-up procedure.* Pick-ups are usually made the same business day depending upon the time of day in which the call is made and the distance that FedEx must travel for the pick-up. Any pick-up that is not possible on the same day of the call is picked up by FedEx no later than the next business day and then promptly forwarded to MI for processing. This allows districts to return all materials immediately upon completion of the test administration. MI encourages districts to return materials as quickly as possible so that processing can begin promptly.

*Login procedure.* MI currently has a system in place that allows log in of all materials within 24 hours of receipt. Upon arrival at MI, all boxes are scanned into our tracking system database where they are logged in and checked against the tracking numbers that are pre-assigned to each district. This provides immediate information on the number of boxes received and their points of origin. Once the login of materials is complete, processing of materials begins at multiple workstations in an effort to meet or exceed the 72-hour requirement for scanning preparation.

Boxes containing non-scannable materials are examined to remove any scannable materials that may have been mixed in error. A separate or “redundancy” check is performed on each box by a second individual at this time to assure that all scannable materials from a particular district are processed together. Any materials located during these searches are placed immediately into the appropriate tote boxes according to the procedure outlined for other scannable materials. The tote boxes of used answer documents are then forwarded to our IT department for scanning and processing, while the boxes of non-scannable materials are held until all scannable materials are processed. The non-scannable boxes are retrieved as soon as possible, but no later than the completion of scannable material processing. The secure materials from those boxes are counted electronically and documented in order to provide information regarding the

quantities of secure materials returned. For the 2012 administration, the security check-in process was completed seven weeks after receiving all materials.

*Secure material check-in.* MI performs a full security check-in using our Security Barcode Check-in Application (SBCA), to capture the booklet barcode number for each booklet returned. This process is labor intensive, but it provides a reliable method of capturing booklet barcode numbers upon return. During the full security check-in, test booklets are unpacked and then scanned at a workstation equipped with a barcode reader and a PC. The barcode of the box into which the booklets are stored, is linked to each set of scanned booklets. Test booklets are stored in boxes of a standard size used for the entire project. All of the barcodes scanned in each box are checked in the master database against the barcodes expected from that district. Any discrepancies are noted and a Security Report is generated, as required. This report is used to inform districts of any secure materials that have not been returned to MI. For Fall 2012 testing, secure and non-secure materials shipped and returned are reported in Tables 3.3.1 through 3.3.3.

**Table 3.3.1**  
**Total Secure Items Shipped in Fall 2012 Administration**

	<b>Mathematics</b>	<b>Reading Day 1</b>	<b>Reading Day 2</b>	<b>Science</b>	<b>Social Studies</b>	<b>Writing Day 1</b>	<b>Writing Day 2</b>
3	155971	140367	140236				
4	156113	139295	139192			139539	139419
5	158690	140787	140730	162208			
6	160543	141782	141671		147969		
7	160015	141662	141546			141998	141913
8	160409	142331	142271	163104			
9					155056		
<b>Total</b>	<b>951741</b>	<b>846224</b>	<b>845646</b>	<b>325312</b>	<b>303025</b>	<b>281537</b>	<b>281332</b>

**Table 3.3.2**  
**Total Secure Items Returned after The End of Testing Windows  
in Fall 2012 Administration**

	<b>Mathematics</b>	<b>Reading Day 1</b>	<b>Reading Day 2</b>	<b>Science</b>	<b>Social Studies</b>	<b>Writing Day 1</b>	<b>Writing Day 2</b>
3	155457	139974	139781				
4	155013	138293	138117			138598	138538
5	157671	139770	139818	161269			
6	159806	141168	140939		147090		
7	159411	141132	141142			141552	141398
8	159606	141824	141726	162490			
9					154056		
<b>Total</b>	<b>946964</b>	<b>842161</b>	<b>841523</b>	<b>323759</b>	<b>301146</b>	<b>280150</b>	<b>279936</b>

**Table 3.3.3**  
**Total Secure Items Not Returned in Fall 2012 Administration**

	<b>Mathematics</b>	<b>Reading Day 1</b>	<b>Reading Day 2</b>	<b>Science</b>	<b>Social Studies</b>	<b>Writing Day 1</b>	<b>Writing Day 2</b>
3	514	393	455				
4	1100	1002	1075			941	881
5	1019	1017	912	939			
6	737	614	732		879		
7	604	530	404			446	515
8	803	507	545	614			
9					1000		
<b>Total</b>	<b>4777</b>	<b>4063</b>	<b>4123</b>	<b>1553</b>	<b>1879</b>	<b>1387</b>	<b>1396</b>

## **CHAPTER 4:**

### **TECHNICAL ANALYSES OF POST-ADMINISTRATION PROCESSING**

#### **4.1. Scanning Accuracy and Reliability**

Measurement Incorporated (MI) has extensive and proven procedures to ensure that we accurately and reliably scan answer documents and collect student responses in preparation for scoring. Our strict handling procedures make sure that each and every answer document is accounted for, tracked, and controlled through every phase of the scanning process. All scanning and scoring applications are fully tested and reviewed using structured testing methodologies before and continually monitored throughout live test materials processing. Any questionable scanned data is flagged for review and correction; thus producing only the highest quality results for reporting.

*Tracking documents.* MI has an application called ObjectTracker to track the location of a scan bin (batch) and its contents (header sheets and test/answer books) throughout processing. Matched batch-tracking barcode labels are affixed to a scan bin and its respective batch-tracking sheet, located on top of the headers and test/answer documents in the scan bin. The batch-tracking barcode is recorded in the ObjectTracker database, which allows us to identify specific scan bins associated with a given school/district and to determine its current status. Because the scan bin ObjectTracker barcodes are carefully scanned in and out of each processing area, it is easy to determine which department is currently in possession of the material. The ObjectTracker application verifies that all batches are accounted for and notifies MI if one is delayed at any particular processing area.

*Cutting multi-page documents.* Scan bins are first forwarded to the cutting area, where cutting personnel remove one scan bin at a time from the cart. The documents are cut using one of MI's four Challenge paper cutters. The cutting operation converts the multi-page answer document into a stack of single sheets ready for scanning. The weight and BTS barcode of the scan bin are recorded in the ObjectTracker database at key points along the processing chain to maintain the integrity of the batch and ensure all documents retain their association with a specific batch.

*Ensuring document integrity.* When scannable materials are printed, each sheet has a scannable and human-readable lithocode value unique to that document. In the unlikely event that a scan bin is dropped at the cutting or pre-scanning stage, the unique lithocode allows answer document integrity to be verified at the scanner as well as when the data is transferred into the project database. Software validations at the scanner ensure that all pages of each student's answer document are accounted for and contain the same lithocode; thus, any pages that are out of order can be easily corrected prior to any other processing.

*Scanner verification and calibration.* Scanning applications that include every scannable document are written using our Virtual Scoring Center™ (VSC™) document setup

application. Each application is tested to ensure that the data derived from all grids appearing on the scannable document are: included in the export file, are accurately read, and return the correct value. A quality control sample of answer documents (test deck) are created so that all possible responses are verified. This structured method of testing provides exact test parameters and a methodical way of determining that the output received from the scanner(s) is correct. The documents and the data file created from them are carefully compared to further ensure that results from the scanner are accurate according to the reporting rules provided by the Bureau of Assessment and Accountability (BAA) staff. Scanner calibration is verified each test cycle prior to the start of scanning, and scanners are recalibrated to specifications prior to each staff shift change so calibration remains constant and precise. In addition, calibration sheets are included in every scan batch to immediately detect scanning problems before a problem can affect subsequent scan batches.

*Image scanning.* The answer documents are scanned in the order they are received and all pages of each complete document are scanned at the same time using our eleven BancTec® IntelliScan® XDS color image scanners. The BancTec® IntelliScan® XDS scanner features a completely open paper path to dramatically improve document throughput. This paper path reduces the time to recover from paper jams and other complications that are common for scanners with more restrictive paper paths. Both sonic and vacuum double-sheet detection technology ensure that every sheet is scanned, allowing reliable interspersed scanning of multi-sized documents. In addition, BancTec has designed custom document integrity software for MI. This application detects out-of-sequence pages allowing operator correction before imaging, thus eliminating post scanning corrective action. The production control technician also weighs the batch when scanning is complete to verify that all the documents in the batch have been scanned. If any discrepancies are detected, the scanner operator submits the batch to a scanning supervisor who investigates and resolves the discrepancy.

To ensure that all sheets in the scan bin are scanned, the last sheet in every bin is an “End of Batch” sheet. If the End of Batch record does not appear in the data file that is imported into the MI database, an error alert is generated, and the technician makes a visual check of the scan bin. The data file is opened again, if necessary, and any missing sheet(s) are appended to the file creating a complete data file.

*Data capture and validation.* The scanning application saves the image data and corresponding index to our Storage Area Network (SAN) that provides fast, secure access to the images. As scan bins are scanned, image files and corresponding index files are created. Once a scan bin is completely scanned, the image and index files are imported from their locations on the SAN into the data capture side of VSC. VSC processes the images using master templates and creates a digital data file from the bubbled information on each page. The Batch Editing operation then uses the images to allow the batch-editing technician to resolve any data integrity issues including lithocode errors, or any image quality issues.

Using the procedures developed by MI and BAA, MI combines the information from the various sources of data (headers and gridded information on answer documents). These multiple, redundant sources of information allow MI to detect discrepancies and ensure that each student is associated with the correct school, grade, and form. MI enforces data validation rules at each stage of processing to reduce last minute data clean-up and ensure the data is accurate, problem-free, and ready for reporting.

*Secure material processing.* After all used answer documents are scanned, the secure, unused third grade answer documents are scanned. This is part of our secure materials processing so that we retain an electronic record of all documents. If documents are located that contain live data, the documents are retrieved and put in scan bins for normal processing as live documents.

*Data correction.* Once all of the information is combined to create the student records, GPA executes data validation routines created specifically for the MEAP. These routines analyze the data and create error tables for answer documents containing questionable data. Common error detection routines include checks for the following situations:

- Inconsistencies in school, grade, or form
- Inconsistencies in headers and answer documents
- Duplicate student barcodes within the same bin or another bin of answer documents
- Duplicate lithocodes
- Missing student barcodes
- Missing or incomplete demographics (such as a blank name)
- Double marks in the demographic and/or multiple-choice grids

MI utilizes a double data correction process to achieve the highest level of quality and accuracy in MEAP student data. Data correction operators use our sophisticated data correction application that retrieves flagged data records and highlights the problem field on a computer screen so it can be resolved. The operator compares the highlighted data to the scanned image of the answer document, and makes any necessary correction. Once an operator corrects a flagged record, the same flagged record is routed to a second data correction operator who repeats the data correction process. After a flagged record is edited by two operators, the data correction application checks that both operators have made identical corrections. In the event that two corrections differ, the record is routed to a supervisory staff member for a third and final resolution. This process continues until all flagged records are examined.

*Test decks and customer acceptance tests.* Test decks and Customer Acceptance Tests are used to verify that the scanning, scoring, and reporting processes are fully functional. First, requirements documents are developed to fully describe scanning, data correction, scoring, and reporting of data. Then test decks and Customer Acceptance Tests (CAT) are created based on those requirements. The test deck process has a very comprehensive set of rules, covering all required scenarios, which are applied to all appropriate grades and content areas. The test deck rules include specifics for handling multiple-answer

documents and constructed response scores for a single student within a single content area, and for aggregating that data at the school, district, ISD, and state levels.

Each BAA assessment test deck begins with answer documents that have been bubbled in order to meet every requirement defined by the BAA as well as specific circumstances defined by MI. Then, more documents are created to represent all logical combinations of requirements and data variations. There is at least one test case for each scenario; each test case requires that we either validate the data that is being captured at scanning or manipulate the data correctly (calculations, overrides, etc.) to yield the appropriate results at the end of the process. In addition, some scenarios have multiple test cases and there are some scenarios that, although not necessary to validate the software functionality, are necessary to provide BAA with scenarios for their own special analysis of particular assessment situations (such as Tested Roster).

The CAT process is divided into multiple stages. Each stage builds on the previous; therefore, BAA must approve results of one stage before MI can perform the tasks associated with the next. The two stages that relate to scanning are discussed below:

- The purpose of the first CAT (Scanning) is to confirm the accuracy of the scan data and that images are captured correctly for all document types. Hardcopies and images of the test deck documents are provided to BAA as well as database tables.
- The second CAT (Data Correction) verifies the accuracy of the data validation and entry systems, specifically that all invalid values in scanned data are sent to Data Correction, and that values entered during Data Correction are transferred accurately to the MI MEAP database.

#### **4.2. Multiple-Choice Scoring Accuracy**

After scanning and data editing, MI scores OMR multiple-choice data using scoring keys. MI uses multiple reviews for accuracy of scoring keys performed by independent staff:

- Analyze the item responses of the students when a significant number of students have been scored in the system.
- Produce a report that allows psychometrician and our personnel to match and review the keys against the percent of correct responses by item on each form and content area. *Appendix G* contains an example of the key check report.
- Send same key check report to the BAA for review.
- Each scored item is marked indicating what the answer was and what the scoring key indicates as the correct answer.
- Questionable data is printed on an edit listing for resolution.
- Each question listed by the edit program is individually reviewed to determine if better or more accurate information can be obtained.

### **4. 3. Erasure Analysis**

The Erasure Analysis is performed on all operational multiple-choice responses once all scanning, data correction, and multiple-choice scoring are complete. Data for each student multiple-choice response is programmatically analyzed to determine if the response contains a mark that exceeds the mark threshold and if the lighter marks are potential erasures. Statistics are captured and aggregated at a school and district level to determine whether the school/district data is outside the state norm. Final results are provided to BAA for review and analysis.

#### **Mark Identification**

The Virtual Scoring Center™ (VSC®) Capture program processes a JPEG grayscale image and assigns a Hex value for each multiple-choice bubble. The Hex range is from 0 to 15; where Hex 0 is the lightest and represents no shading contained in the bubble and Hex 15 is the darkest and represents a dark, student filled bubble. A student selected response is captured when the Hex value for the bubble is Hex 12 (definite mark threshold) or above. A bubble detected in the range of 9 to 11 is captured as the student response if no other bubble for the multiple-choice question is above an 8. A bubble is considered an erasure if the Hex value for the bubble is greater than 5 and less than 12 and not identified as the student response.

#### **Erasure Identification**

Using the VSC Capture image processed Hex value for each bubble in a multiple-choice questions, each Hex value is analyzed to determine if an erasure is present. A flag is set for each bubble that is detected as an erasure. The iErasureA flag is set if the A bubble was erased, iErasureB is set if the B bubble was erased, and so on.

#### **Erasure Analysis**

The answer key for each test is used to compare the student selected response, the correct answer, and the erased bubble to determine multiple-choice erasure results. There are three results for an erased multiple-choice question: wrong answer to correct answer; correct answer to wrong answer; or wrong answer to wrong answer. A result flag is set for each erased multiple-choice case.

### **4.4. Results of Constructed Response Scoring Procedures**

The MEAP 3-9 statewide assessment includes measures in which the examinees must construct their own response for some of the questions. For example, examinees may be required to provide a two or three sentence response to a reading comprehension item, or demonstrate the appropriate use of a geometry formula. The procedure for scoring these responses is provided.

Outlined below is the scoring process Measurement Incorporated follows. This procedure is used to score responses to all MEAP constructed response or written composition items.

#### **4.4.1. Rangefinding and Rubric Review**

MI project leadership personnel have worked successfully with rangefinding committees from many states over the last three decades, and we conducted many successful rangefinding meetings during our previous contracts with the State of Michigan. We support this important element of the scoring process and will conduct both the Initial and Final rangefinding meetings, in coordination with BAA staff, for each administration of MEAP. Rangefinding committees are convened for each grade level and content area of the assessment; for MEAP ELA, there are separate committees for the reading and writing tasks at each grade level.

MI understands that each rangefinding committee consists of BAA staff, 5 to 8 Michigan teachers of the appropriate grade level (based on the answer to question #30) , and at least one project monitor or scoring director from MI. We currently have a number of MI project monitors available and scoring directors with experience in conducting rangefinding meetings for MEAP under previous contracts, and we added additional leadership staff that has previous experience in conducting rangefinding for other states. The meetings for each subject and grade level committee last for no longer than three days.

MI recruits and trains active Michigan teachers to be members of rangefinding committees for the subject and grade in which they teach. We use BAA lists of potential rangefinding committee members to recruit participants. We train these teachers so they are prepared to apply scoring criteria to the constructed-response items being tested.

MI scoring staff facilitates the rangefinding meetings with several goals in mind. First, to accurately and consistently apply the rubric to each student response and to ensure the rubrics are viable. MI staff keeps careful records of all scoring decisions made at these meetings, including notes about which student responses are problematic to score and which are not. We document and archive the records of all of the decisions and provide them to the BAA.

Another goal is to ensure that the participating Michigan educators feel confident about the rangefinding process and the resulting scoring decisions. Many Michigan educators inform us that rangefinding is their favorite of all professional development activities; this feedback gave us much satisfaction during our work on previous Michigan assessments

MI scoring directors have extensive experience in rangefinding situations and after rangefinding these same scoring directors construct the training materials and conduct the training of the readers. This continuity of leadership helps to ensure that the

scoring criteria established at the rangefinding meetings are applied during training and scoring.

**Initial rangefinding.** The Initial rangefinding meeting occurs before the handscoring of the field tested constructed-response items. This round of rangefinding is to verify that all score points are represented and that the rubric is viable. Items may be discarded (or more likely, sent back to the development contractor for editing) at this stage. MI understands that Writing has pre-established rubrics while Reading has rubrics specific to the items themselves.

In order to provide a variety of responses and potential score points, MI leadership staff selects and copies representative field test responses in advance. We carefully select approximately 25 sample responses per item. MI brings sufficient copies of the rangefinding responses to the meetings so that each participant has his/her own set. In addition, we provide all necessary supplies. We use previous MEAP operational training materials to train the members of the rangefinding committees.

Before any secure materials are distributed, all committee members are required to sign a confidentiality/nondisclosure form. MI keeps these forms on file for the duration of the contract. During the meetings, each committee member has a complete set of rangefinding materials with which to work. These materials are numbered so that we can account for them all at the end of the meetings. Each committee member is permitted to take notes on these sets. During the meetings, the committee members have access to all materials related to a particular item until no more discussion of that item is required. When discussion of an item is completed, all responses, rubrics, and item sheets associated with that item are collected before a new item is distributed. At the end of the rangefinding meetings, all materials are recycled under secure conditions.

Final versions of the scoring rubrics for the test items are produced based on the decisions made by the rangefinding committees. The revised rubrics and responses that were scored at the Initial rangefinding meeting are used to train the readers to score the field test sample. Readers for the field test scoring have previously qualified for and have scored the operational test items and have demonstrated a high degree of reliability. They are able to accurately apply the scoring criteria for the field test items after being trained with the rangefinding responses. In addition, during the scoring of the field test items, any new scoring questions are addressed by communication between the BAA Staff and MI scoring leadership.

**Final rangefinding.** After the operational items have been selected for each assessment, MI conducts the Final rangefinding meetings.

Before the meetings, MI leadership staff selects and copies field test responses that represent all score points for each item as well as some unusual responses that were referred to the BAA for decisions during the field test scoring. We select 125 field-tested responses for each item. MI assembles, copies, and brings these materials to

range-finding along with any supplies needed to conduct the meetings. All of the procedures for security instituted at the Initial range-finding meetings are replicated at the Final meetings. We train the Final range-finding committee members with the anchor responses and the rubrics that were finalized at the Initial range-finding meetings.

The Final range-finding meetings are conducted to establish “true” scores for a representative sample of student responses to the writing prompts and constructed-response tasks in each content area and at each grade level. We use feedback from the educators who participate in these meetings to select exemplar responses that are used to construct the materials for reader training (anchor sets, training/qualifying sets, validity/calibration sets). If any responses selected are not acceptable to the range-finding committee or to the BAA content specialists, or if the numbers of responses are not sufficient to construct the reader training materials, MI provides additional responses for review and approval by BAA staff.

At the conclusion of each meeting, there is a final step in the range-finding process. After the range-finding responses have been discussed and have received a final score, the range-finding committee sorts their responses into stacks by score point and rereads the responses at each score point to ensure consistency. MI scoring directors perform an additional check for consistency after the meetings are over. They read the sorted responses again and confer with the BAA Office and the MI Handscoring Manager if there are problems with consistency.

MI’s considerable experience in range-finding situations is brought to the MEAP assessments to ensure the continuation of solid criteria and consistent scores. We take great care to work with the BAA and Michigan educators in developing guidelines which promote consistency in scoring in future years.

#### **4.4.2. Rater Selection**

MI maintains a large pool of qualified, experienced readers at each scoring center. We need only inform them that a project is pending and invite them to return. MI routinely maintains supervisors’ evaluations and performance data for each person who works on each scoring project in order to determine employment eligibility for future projects. We employ many of our experienced readers for this project and recruit new ones as well.

MI procedures for selecting new readers are very thorough. After advertising and receiving applications, MI staff review the applications and schedule interviews for qualified applicants. Qualified applicants are those with a four-year college degree, preferably in English, language arts, education, or a related field. Each qualified applicant must pass an interview by experienced MI staff, write an acceptable essay, and receive good recommendations from references. We then review all the information about an applicant before offering employment.

In selecting team leaders, MI's management staff and scoring directors review the files of all returning staff. They look for people who are experienced team leaders with a record of good performance on previous projects and also consider readers who have been recommended for promotion to the team leader position.

MI is an equal opportunity employer that actively recruits minority staff. Historically, our temporary staff on major projects averages about 51% female, 49% male, 76% Caucasian and 24% minority. MI strongly opposes illegal discrimination against any employee or applicant for employment with respect to hiring, tenure, terms, conditions, or privileges of employment, or any matter directly or indirectly related to employment, because of race, color, religion, sex, age, handicap, national origin, or ancestry.

MI requires all handscoring project staff (scoring directors, team leaders, readers, and clerical staff) to sign a Confidentiality/Nondisclosure Agreement before receiving any training or other secure project materials. The employment agreement indicates that no participant in training and/or scoring may reveal information about the test, the scoring criteria, or scoring methods to any person. A copy of this agreement is enclosed in Section 6: Work Samples.

#### **4.4.3. Rater Training**

All readers hired for MEAP handscoring are trained using the rubric(s) approved by the BAA and responses selected during the rangefinding meetings. Readers are placed into a scoring group that corresponds to the subject that he/she has taught or studied. Within each group, readers are divided into teams consisting of one team leader and 10-15 readers. Each team leader and reader are assigned a unique number for easy identification of their scoring work throughout the scoring session.

After the contracts and nondisclosure forms are signed, and the introductory remarks are given by the scoring director, training begins. Reader training and team leader training follow the same format, except that team leaders are required to annotate each response in the training sets, while readers are encouraged to take notes. The scoring director presents the writing or constructed-response task and introduces the scoring guide (anchor set), then discusses, roomwide, each score point. This presentation is followed by practice scoring on the training/qualifying sets. The scoring director reminds the readers to compare each training/qualifying set response to anchor responses in the scoring guide to assure consistency in scoring the training/qualifying responses.

Each reader has a clean copy of the training/qualifying sets as well as a score sheet on which to record training set scores. Once the readers score these responses, they take their score sheets to their team leader. The team leader will record the percentage of correct scores both on the reader's score sheet and on a logbook that is kept to record performance of all team members on all training/qualifying sets. This function is also performed by scoring directors during team leader training. The team leaders' log

books are submitted to the BAA Office. If an BAA representative is on-site during team leader and/or reader training, the representative has access to these documents as each set is completed.

Because it is easy in a large group to overlook a shy reader who may be having difficulty, readers break into teams to discuss the responses in the training/qualifying sets. This arrangement gives readers an opportunity to discuss any possible points of confusion or problems in understanding the criteria. The scoring director will also “float” from team to team, listening to the team leaders’ explanations and adding additional information when necessary. If a particular response or type of response seems to be causing difficulty across teams, the scoring director discusses the problem roomwide to ensure that everyone hears the same explanation. Once each team has finished discussing the first set, the readers score the next set. Training continues until all training/qualifying sets have been scored and discussed.

Like team leaders, readers must demonstrate their ability to score accurately by attaining the qualifying agreement percentage established by the BAA before they may read actual student responses. Any readers unable to meet the standards set by the BAA are dismissed. All readers understand this stipulation when they are hired. MI is always sensitive to the need for accurate and consistent scoring, and any team leader or reader who is not able to demonstrate both accurate and consistent results during training is paid for time spent and dismissed.

Training is carefully orchestrated so that readers understand how to apply the rubric in scoring the responses, learn how to reference the scoring guide, develop the flexibility needed to deal with a variety of responses, and retain the consistency needed to score all responses accurately. In addition to completing all of the initial training and qualifying, a significant amount of time is allotted for demonstrations of the VSC handscoring system, explanations of how to “flag” unusual responses for review by the scoring director, and instructions about other procedures which are necessary for the conduct of a smooth project.

Reader training requires approximately three days (actual time varies by grade and content area). Readers generally work 7.0 hours per day, excluding breaks. Evening shift readers work 4.5 hours, excluding breaks. A typical reading schedule is shown below.

Day Shift

8:15 a.m. – 10:00 a.m.	Reading
10:00 a.m. – 10:15 a.m.	Paid Break
10:15 a.m. – 12:00 p.m.	Reading
12:00 p.m. – 12:45 p.m.	Lunch
12:45 p.m. – 2:00 p.m.	Reading
2:00 p.m. – 2:15 p.m.	Paid Break
2:15 p.m. – 4:00 p.m.	Reading

#### Evening Shift

5:00 p.m. - 6:45 p.m.	Reading
6:45 p.m. - 7:15 p.m.	Dinner
7:15 p.m. - 9:00 p.m.	Reading
9:00 p.m. - 9:15 p.m.	Paid Break
9:15 p.m. - 10:15 p.m.	Reading

### **4.5. Rater Statistics and Analyses**

#### **4.5.1. Calibration**

A concern regarding the scoring of any open-response assessment is the reliability and accuracy of the scoring. MI appreciates and shares this concern and continually develops new and technically sound methods of monitoring reliability. Reliable scoring starts with good rangefinding meetings, development of detailed scoring rubrics and training materials, and thorough training sessions by experienced trainers. Quality results are achieved by daily monitoring of each reader. Unbiased scoring is ensured because the only identifying information on the student response is the identification number. Unless the students sign their names, write about their hometowns, or in some way provide other identifying information, the readers have no knowledge of them.

In addition to extensive experience in the preparation of training materials and employing management and staff with unparalleled expertise in the field of handscored educational assessment, MI constantly monitors the quality of each reader's work throughout every project. Methods that are used to monitor readers' scoring habits during Michigan handscoring projects include use of the following:

*Reader Status Reports.* MI has developed and operates a comprehensive system for collecting and analyzing scoring data. After the readers' scores are submitted into the VSC handscoring system, the data is uploaded into the primary Project Command Center (PCC) servers located at our corporate headquarters in Durham, North Carolina. These scores are then validated and processed according to the specifications set out by the BAA.

There are currently more than 20 reports available that can be customized to meet the information needs of BAA and MI's scoring department. We provide BAA with reports that include the following data:

- Reader ID and team
- Number of responses scored
- Number of responses assigned each score point (1-4 or other)
- Percentage of responses scored that day in exact agreement with a second reader
- Percentage of responses scored that day within one point agreement with a second reader
- Number and percentage of responses receiving adjacent scores at each line (1/2,

2/3, 3/4)

- Number and percentage of responses receiving nonadjacent scores at each line
- Overall reliability index (taken from exact agreements and one point discrepancies with other readers)
- Number of correctly assigned scores on the validity responses

Updated “real-time” reports are available that show both daily and cumulative (project-to-date) data. These reports are available for access by BAA staff via a secure website. Our reporting system provides 24-hour on-line access to the reader status reports through the use of a user name and password that is provided to the BAA at the beginning of each test administration. This allows any BAA staff member with access to review a scoring report whenever they prefer. We discuss quality control procedures and reporting at the Kick-Off meetings and the BAA lets us know if they want reports more comprehensive than those listed above. Sample reliability reports are included in Section 6.

The handscoring project monitors at each MI scoring center also have access to the PCC system and they provide updated reports to the scoring directors several times a day. MI scoring directors are experienced in examining these reports and using the information to determine the need for retraining of individual readers or the group as a whole. It can easily be determined if a reader is consistently scoring “too high” or “too low,” as well as the specific score points with which they may be having difficulty. The scoring directors share such information with the team leaders and direct all retraining efforts.

#### **4.5.2. Rater Monitoring and Retraining**

Team leaders spot-check (read behind) each reader’s scoring to ensure that he/she is on target, and conduct one-on-one retraining sessions about any problems found. At the beginning of the project, team leaders read behind every reader every day; we become more selective about the frequency and number of read-behinds as readers become more proficient at scoring. The Daily Reader Reliability reports and validity/calibration results are used to identify the readers who need more frequent monitoring.

Retraining is an ongoing process once scoring is underway. Daily analysis of the Reader Status Reports alert management personnel to individual or group retraining needs. If it becomes apparent that a whole team or a whole group is having difficulty with a particular type of response, large group training sessions are conducted. Standard retraining procedures include room-wide discussions led by the scoring director, team discussions conducted by team leaders, and one-on-one discussions with individual readers. It is standard practice to conduct morning room-wide retraining at MI each day, with a more extensive retraining on Monday mornings in order to re-anchor the readers after a weekend away from scoring.

Our quality assurance/reliability procedures allow our handscoring staff to identify

struggling readers very early and begin retraining immediately. During the time when we retrain these readers, we also monitor their scoring intensively to ensure that all responses are scored accurately. In fact, the monitoring we do is also used as a retraining method (we show readers responses that they have scored incorrectly, explain the correct scores, and have them change the scores). Our retraining methods are very successful in helping readers to become accurate scorers.

#### **4.5.3. Rater Dismissal**

When read-behinds or daily statistics identify a reader who is unable to maintain acceptable agreement rates, the reader is retrained and monitored by scoring leadership personnel. A reader may be released from the project if retraining is unsuccessful. In these situations, all items scored by a reader during the timeframe in question can be identified, reset, and released back into the scoring pool. The aberrant reader's scores are deleted, and the responses are redistributed to other qualified readers for rescoring.

#### **4.5.4. Score Resolution**

Each student response is scored holistically by a trained and qualified reader using the scoring scales developed and approved by the BAA. There will be a blind 20% second read for reliability purposes.

#### **4.5.5. Inter-Rater Reliability Results**

Inter-rater agreement is expressed in terms of exact agreement (Reader Number One's score equals Reader Number Two's score) plus adjacent agreement (+/-1 point difference). Summary statistics of inter-rater agreement and rater validity agreement for all the constructed-response items are presented in Appendix Q.

Inter-rater agreement as expressed by sum of perfect and adjacent percent of agreement ranges from 98.0% to 99.1% for Reading and 97.0% to 99.7% for Writing.

#### **4.5.6. Rater Validity Checks**

Scoring directors select responses from rangefinding which are loaded into the VSC system as validity responses. The "true" or rangefinding scores for these responses are entered into a validity database. These responses are sent out into the rooms each day to be scored by the readers. A validity report is generated that includes the response identification number, the score(s) assigned by the readers, and the "true" scores established by the rangefinding committee. A daily and project-to-date summary of percentages of correct scores and low/high considerations at each score point is also provided.

## **CHAPTER 5: MEAP REPORTS**

### **5.1. Description of Scores**

#### **5.1.1. Scale Score**

Scale scores are statistical conversions of raw scores that adjust for slight differences in underlying ability levels at each score point and permit comparison of assessment results across different test administrations within a particular grade and subject. Each year new test forms are developed. These new forms never contain exactly the same questions as the previous forms. In order to have a fair comparison across time, it is necessary to have a scale score that means the same across different administrations. On MEAP grades 3-9 assessment, scale score with a standard deviation of 25 is developed, and a score of X00 is assigned to a student of grade X who barely *meets* Michigan standards. For example, a score of 500 is assigned to a grade 5 student who barely meets Michigan standards. Scale scores are not comparable across grade levels. A scale score of 500 on the grade 4 assessment does not indicate that the fourth-grade student would be considered as meeting standards on the grade 5 assessment. Details of the development of MEAP scale scores are described in Chapter 7, section 7.2. The scale score is stable because it allows for students' scores to be reported on the same scale regardless of which year they took the assessment, and which form of the assessment the student took.

Schools can use scale scores to compare the achievement of groups of students across years. These comparisons can then be used to assess the impact of changes or differences in instruction or curriculum. The scale score can be used to determine whether students are demonstrating the same skill and ability across cohorts within a grade and subject.

#### **5.1.2. Raw Score**

In addition to scale scores, sub-content raw scores are reported in the score reports. These scores are the sum of raw points earned as classified into specific content categories. Total raw scores are also reported. Several values that are derived from the raw scores are added to assist in interpreting the raw scores: maximum possible score points, percent correct and aggregate averages (for school and district level reports).

#### **5.1.3. Performance Level**

To help parents and schools interpret the reported score values, performance levels are identified. A performance level is a range on the score scale that corresponds to student achievement levels. The MEAP student performance levels are:

1. *Advanced*;
2. *Proficient*;
3. *Partially Proficient*; and
4. *Not Proficient*.

The divisions between the levels are called cut scores. The cut scores are recommended by a panel comprised of educators and other stakeholders throughout the state. This panel uses detailed descriptions of what students in each of the performance descriptions should know and be able to do. Based upon these detailed descriptions and actual assessment items, the panel recommends the score that best separates each performance level from the next. See Chapter 6 for detailed standard setting process. The Michigan State Board of Education approves the final cut scores and performance level ranges.

#### 5.1.4. Mini-Categories

Mini-categories are subcategories that are created within each performance level for Mathematics and Reading that are used by BAA as part of their growth model. There are nine mini-categories on each assessment; Not Proficient low, Not Proficient Mid, Not Proficient High, Partially Proficient Low, Partially Proficient High, Proficient Low, Proficient Mid, Proficient High, and Advanced. These mini-categories are created by considering the performance levels determined from the cut scores and considering the conditional standard error of measurement on each assessment. Mini-categories are created to subdivide the performance levels such that a change in achievement from one mini-category to the next exceeds the conditional standard error of measurement. This is done to ensure that moving from one mini-category to the next will represent a change in achievement that is not purely due to measurement error (Martineau 2007; Wyse, Zeng, & Martineau, 2011). The mini-categories are reported for fall 2011 and will be used to determine performance level change in Michigan's growth model in 2012.

#### 5.1.5. Performance Level Change

Performance level changes will be reported in 2012 for MEAP and it is determined through a comparison of the mini-category that a student obtained the current year in relationship to the mini-category that they obtained in the previous year. Table 5.1 displays how these scores are used in Michigan's transition table growth model (Martineau, 2007).

Table 5.1.5: MEAP Transition Table Growth Model

Year X Grade Y MEAP Performance Level		Year X+1 Grade Y+1 MEAP Performance Level								
		Not Proficient			Partially Proficient		Proficient			Advanced
		Low	Mid	High	Low	High	Low	Mid	High	Mid
Not Proficient	Low	M	I	I	SI	SI	SI	SI	SI	SI
	Mid	D	M	I	I	SI	SI	SI	SI	SI
	High	D	D	M	I	I	SI	SI	SI	SI
Partially Proficient	Low	SD	D	D	M	I	I	SI	SI	SI
	High	SD	SD	D	D	M	I	I	SI	SI
Proficient	Low	SD	SD	SD	D	D	M	I	I	SI
	Mid	SD	SD	SD	SD	D	D	M	I	I
	High	SD	SD	SD	SD	SD	D	D	M	I
Advanced	Mid	SD	SD	SD	SD	SD	SD	D	D	M

In Table 5.1.5, one can see that there are five different types of transitions that a student can take from one year to the next. The five different transitions are shown with five different colors and include a significant decline (SD), decline (D), maintain (M), improve (I), or significant improve (SI). A student significantly declines if their performance goes up by more than two mini-categories, a student declines if their performance goes down by one or two mini-categories, a student maintains if their performance is in the same mini-category in consecutive years, a student improves if their performance goes up by one or two mini-categories, a student significantly improves if their performance goes up by more than two mini-categories. Students that are in the improve or significant improve categories will be considered proficient due to growth for AYP purposes in 2012 and moving forward.

## 5.2. Scores Reported

Brief descriptions of MEAP score reports are provided below. More extensive descriptions with samples are included in the *Guide to Reports, Grades 3–9* (Fall 2012).

- *Summary Report* is a comparative set of mean scale score information for each grade level, summarized by school, district, ISD, and state. All content areas and levels of performance are reported. The report is generated for three student populations: All students; students with disabilities (SWD); and all except students with disabilities (AESWD).
- *Demographic Report* provides a summary breakdown of scores by demographic subgroup for each content area assessed. Summary data reported includes the number of students assessed in each subgroup, the mean scale score, the percentage of students attaining each performance level, and the percentage of students that met or exceeded Michigan standards within each content area. The Demographic Report is generated for all students, SWD, and AESWD. The demographic subgroup scores are aggregated by school, district, ISD, and state. The demographic subgroups reported are gender, ethnicity, economically disadvantaged (ED), English language learners (ELL), formerly limited English proficient (FLEP), and migrant.
- *Feeder School Report* is a summary report provided to feeder schools at transition grade levels. For example, District A has three elementary schools (K-5) feeding into one middle school (6-8). Each elementary school will receive a Feeder School Report summarizing the data for current sixth-grade students that were enrolled in their elementary school at the end of Grade 5.
- *Item Analysis Report* provides summary information for each multiple-choice item, and each constructed response item on the assessment, including the primary Michigan benchmark (GLCE) measured by each item. The summary information reports the percentage of students selecting each response. The report is generated for all students, SWD, and AESWD. The aggregate data is reported by class or group, school, district, and state.

- *Class Roster Report* provides summary score information by class, for each strand and benchmark (GLCE) assessed within each content area, as well as detail information for each student assessed.
- *Individual Student Report (ISR)* provides a detailed description of each student's performance in the content area assessed. It is designed to help educators identify the academic strengths of their students and the areas that may need improvement.
- *Student Record Label* is provided for each student assessed during the Fall 2012 cycle. The labels are mailed to the school for placement in the student record file.
- *Parent Report* provides a summary description of their student's performance in each content area assessed. This report is designed to help parents and guardians identify the academic strengths of their student and areas that may need improvement.

### **5.3. Appropriate Score Uses**

MEAP assessment results have several uses for individual students and for comparing the performance of groups.

#### **5.3.1. Individual Students**

Individual student scale score and performance level can be used to evaluate the student's achievement compared to the standards set by the state. Assessment results yield from MEAP can also be used to compare the performance of an individual student to the performance of a similar demographic or program group or to an entire school or district. For example, the scores for a Hispanic student in a Title I program can be compared to the average scores of Hispanic students, Title I students, all the students in a district, or any combination of these aggregations.

Other scores provide information about academic areas of relative strength or weakness. The scores in strands and benchmarks (GLCE) can help educators identify the academic strengths of their students and the areas that may need improvement. The sub-scores are not as reliable as total test scores and should be used in conjunction with other evaluations of performance to provide a portrait of the student's achievement.

#### **5.3.2. Groups of Students**

Test scores can be used to compare the performance of different demographic or program groups to each other. All scores can be analyzed within the same subject and grade for any single administration to determine which demographic or program group had, for example, the highest average performance, the lowest percent meeting minimum expectations, or the highest percent mastery of the "chance and data" objective on the MEAP mathematics tests.

Other scores can be used to help evaluate academic areas of relative strength or weakness. The sub-score strands provide information to help identify areas where further

diagnosis may be warranted when a school's performance is contrasted with the district's or state's.

Because the MEAP tests are designed to measure content areas within the required standards, considering test results by subject area and by objective may be helpful when evaluating curriculum, instruction and their alignment to standards. Generalizations from test results may be made to the specific content domain represented by the objective or set of objectives being measured on the exam. However, because the tests are measuring a finite set of skills with a limited set of items, generalizations should be made only to student achievement as measured on a particular test. All instruction and program evaluations should include as much information as possible to provide a more complete picture of performance.

In addition, all test scores can be compared to regional and statewide performance within the same subject area and grade for any administration.

### **5.3.3. Item Statistics**

*Appendix H* provides item level statistics by grade and subject area for the MEAP. These statistics represent the item characteristics used most often to determine how a group of students performed on a particular item and whether an item functioned in an appropriate manner. The item mean is synonymous with the item *p*-value. For multiple-choice items it is the percent of all students that responded to an item correctly. For constructed-response items it is the average of score points earned on the item. Adjusted mean can be used to compare the difficulty of dichotomous and polytomous items and it is computed by the formula below:

$$Adj\_Mean = \frac{item\_mean - item\_min}{item\_max - item\_min}.$$

For multiple-choice items, the adjusted mean is the same as item mean. Rasch item difficulty and its standard error are also provided in the table. The mean square fit (MNSQ) statistics are used to determine whether items are functioning in a way that is congruent with the assumptions of the Rasch mathematical model. Under these assumptions, how a student will respond to an item depends on the ability of the student and the difficulty of the item, both of which are on the same measurement scale. If an item is as difficult as a student is able, the student will have a 50-50 chance of getting the item correct. If a student is more able than an item is difficult (under the assumptions of the Rasch model), that student has a greater than 50% chance of answering the item correctly. On the other hand, if the item is more difficult than the student is able, he or she has a less than 50% chance of responding correctly. Rasch fit statistics estimate the extent to which an item is functioning in this predicted manner. Items showing a poor fit to the Rasch model typically have values in excess of 1.3. Items may not fit the Rasch model for several reasons, all of which relate to students responding to items in an unexpected way. For example, if an item appears to be easy but consistently solicits an incorrect response from high scoring students, the fit value will likely be over 1.3.

Similarly, if a difficult item is answered correctly by several low ability students, a fit flag may be generated. In most cases the reason behind why students respond in unexpected ways to a particular item is unclear. However, on occasion it is possible to determine the cause of an item's misfit by re-examining the item and its distracters. For example, if several high ability students miss an easy item, re-examination of the item may show that it actually has more than one correct response.

Two types of MNSQ values are presented in *Appendix H*, OUTFIT and INFIT. MNSQ OUTFIT values are sensitive to outlying observations. Consequently, OUTFIT values will be large when students perform unexpectedly on items that are far from their ability level. For example, easy items for which very able students answer incorrectly and difficult items for which less able students answer correctly. MNSQ INFIT values are sensitive to behaviors that affect students' performance on items near their ability estimates. Therefore, high INFIT values would occur if a group of students of similar ability consistently responded incorrectly to an item at or around their estimated ability. For example, under the Rasch model the probability of a student with an ability estimate of 1.00 responding correctly to an item with a difficulty of 1.00 is 50%. If several students at or around the 1.00 ability level consistently miss this item such that only 20% get the item correct, a fit flag will likely be generated because the performance of the item is inconsistent with the expectations of the model. Mis-keyed items or items that contain cues to the correct response (i.e., students get the item correct regardless of their ability) may elicit infit flags. Tricky items, or items that may be interpreted to have double meaning may elicit outfit flags.

Item-total correlations provide another measure of the congruence between the way an item functions and our expectations. Typically we expect students with high ability (i.e., those who perform well on the MEAP overall) to get items correct, and students with low ability (i.e., those who perform poorly on the MEAP overall) to get items incorrect. If these expectations are accurate, the point-biserial (i.e., item-total) correlation between the item and the total test score will be high and positive, indicating that the item is a good discriminator between high ability and low ability students. A correlation value above 0.20 is considered acceptable; values closer to 1.00 indicate greater discrimination. A test comprised of maximally discriminating items will maximize internal consistency reliability. The correlation is a mathematical concept, and therefore not free from misinterpretation. Often when an item is too easy or too difficult, the point-biserial correlation will be artificially deflated. For example, an item with a p-value of .95 may have a correlation of only 0.15. This does not mean that this is a "bad" item. The low correlation is simply a side-effect of the item difficulty. If the item is extremely easy to everyone, not just high scoring students, getting it correct results in a low correlation. Due to these potential misinterpretations of the correlation, it is important to remember that this index alone should not be used to determine the quality of an item. It should instead be used as an indicator to flag items that need further investigation.

#### **5.3.4. Frequency Distributions**

*Appendix W* provides scale score distributions in histograms by grade and subject area for the MEAP assessment. The scale score distributions classified by gender, ethnicity, socioeconomic status, and English proficiency allow us to compare performance of different sub-groups of students in MEAP assessment.

## **CHAPTER 6: PERFORMANCE STANDARDS**

This chapter presents the MEAP proficiency level cut score and their development. The MEAP proficiency level cut score were originally determined through a standard setting in 2005. New set of MEAP proficiency level cut scores was derived from a special study in 2011 and has been applied to the MEAP administration since 2011.

### **6.1. Development of Standard Setting Performance Level Descriptors**

In this section, a brief description of the development of standard setting performance level descriptors (PLDs) is presented. The purpose of the PLDs meetings was to develop specific PLDs within and across grades in each subject area for the Michigan Educational Assessment Program (MEAP). The meetings were convened over a two day period, September 23-24, 2005, following the content alignment study being conducted by Norm Webb. Teachers participating in the content alignment study stayed an extra day to develop the PLDs to be used at the MEAP standard setting in January, 2006. A sample of those teachers stayed an additional half day to align the newly developed grade level PLDs across all grades.

Panelists first were asked to review the general performance level descriptors (PLDs), which were brief two to three sentence descriptions of each level. These PLD highlighted the words and phrases that most distinguished one level from another. The group facilitator led a brief discussion around these distinguishing factors and then reiterated the need to operationalize the descriptors for a given grade. They then reviewed the subject specific PLDs, which were more detailed descriptions at each level. Panelists were given time to review the Assessment Blueprint and grade level content expectations (GLCEs) associated with the grade and subject for which they were writing PLDs.

Panelists were asked to use the general and subject specific PLDs, the assessment blueprint, and GLCEs to define expectations for students at each of the four proficiency levels: Level 1: Exceeded Michigan Standards, Level 2: Met Michigan Standards, Level 3: Basic Level, and Level 4: Apprentice. Panelists were reminded that the goal of the task was to operationalize the performance level descriptors to foster a common understanding of what it meant to be classified within a given level. Individuals contributed to a final list representing the general consensus.

Table 6.1.1 lists the tasks and timeframe of steps in the planning for and completing the specific standard setting performance level descriptors (SS PLDs).

**Table 6.1.1: Tasks and Deliverables for Development of SS PLDs**

<b>Task/Deliverable</b>	<b>Owner</b>	<b>Deadline/ Timeframe</b>
Draft outline of SS PLDs plan	PEM	6/7/05
Feedback/edits to outline	BAA	6/14/05
Finalize dates and location (need content staff to represent MEAP)	BAA	6/30/05
General and subject specific PLDs defined	BAA	7/22/05
Feedback/edits to plan	BAA	7/29/05
Invite members to participate in committees	BAA	8/1/05
SS PLDs development plan/script to BAA	PEM	8/12/05
Feedback/edits to script	BAA	8/18/05
Finalize script	PEM	8/19/05
Finalize list of attendees	BAA	8/30/05
Train facilitators in Iowa City	PEM	9/1/05
Train facilitators in Austin	PEM	9/6/05
Facilitate committees to develop SS PLDs	PEM	9/23-9/24/05
SS PLD write-up	PEM	10/27/05

***Panelists***

For pairwise grade SS PLDs: Panelists were familiar with the grade level and content area. In fact, most panelists taught in the grade level and content area for which they were developing SS PLDs. Since the MEAP was administered in the Fall, it was assessing content from the previous grade. For instance, grade 5 mathematics MEAP was assessing content from fourth grade. Following the content alignment study, the teachers stayed on an extra day (and a half) to write the SS PLDs. Table 6.1.2 lists the committee makeup by subject and grade. For ELA and math, each subject had three committees. Each committee was to be comprised of four teachers writing the SS PLDs: four for grades 3 and 4, four teachers writing the SS PLDs for grades 5 and 6, and four teachers writing the SS PLDs for grades 7 and 8. For science, the committee was to include the eight members from the content alignment study, for both grades 5 and 8 (each with four teachers per grade). Social studies was not undergoing content alignment, so teachers were brought in a half-day early to review the content before starting the SS PLD task, with a makeup similar to the science committees: four teachers for grade 6 and four teachers for grade 9. However, the target set of panelists were not met, and as seen in Table 6.1.2 below, there were fewer panelists than anticipated for each committee. In essence, none of the out-of-state teachers from the content alignment study stayed to help set the SS PLDs, and not all Michigan teachers came to either set of meetings. The only exception was in the science committee, where there was an abundance of Michigan science educators who participated. For each committee, there is a summary of the demographic information collected for each panelist attached to this report.

**Table 6.1.2: Panelist distribution by grade**

<b>Subject</b>	<b>Grade</b>	<b>Target: Panelists</b>	<b>Actual Panelists</b>
ELA	3/4	four teachers: two from Michigan and two from out-of-state	two Michigan educators
ELA	5/6	four teachers: two from Michigan and two from out-of-state	two Michigan educators
ELA	7/8	four teachers: two from Michigan and two from out-of-state	three Michigan educators
Math	3/4	four teachers: two from Michigan and two from out-of-state	four Michigan educators
Math	5/6	four teachers: two from Michigan and two from out-of-state	three Michigan educators
Math	7/8	four teachers: two from Michigan and two from out-of-state	three Michigan educators
Science	5/8	eight teachers: four from Michigan and four from out-of-state	seven Michigan educators
Social Studies	6/9	eight teachers: four from Michigan and four from out-of-state	four Michigan educators

There was also vertical alignment of SS PLDs, which was a second component to the meetings. The vertical alignment component of the SS PLDs brought together all panelists from the grade level committees in math for a total of 10 panelists and in ELA, for a total of seven panelists. The purpose of this super-committee was to align the SS PLDs across grades, in ELA and in math. There was no vertical alignment aspect to the science or social studies groups, since both subjects only had two grades being assessed.

Facilitators were consistent in using the script, agenda, and materials provided. For the math, ELA and science committees, panelists began the SS PLDs meetings at 1:00 pm on September 23, 2005. Since social studies was not undergoing content alignment, the social studies teachers began at 8:00 am to have a half-day to learn about the content alignment terminology as well as to become more intimately involved in understanding the test items and the content standards and benchmarks being measured by the MEAP assessments.

At 1:00 pm, teachers from all four subjects began their participation in the SS PLD meeting. Bureau of Assessment and Accountability (BAA) introduced the need for aligned PLDs across grades as part of the standard setting process. Panelists then reviewed the general PLDs that described the four performance levels and reviewed the activities they would be involved in for the remainder of their time. Panelists were then divided into small groups, based on the grades and subjects they were participating in, for a total of eight separate committees. For ELA, there were three committees: one for grades 3 and 4, one for grades 5 and 6, and one for grades 7 and 8. Similarly, there were three such committees for mathematics. For science, there was one committee for grades 5 and 8, and for social studies there was one committee for grades 6 and 9.

The ELA and math committees began by reviewing the content alignment terminology. This was to insure that the frame of reference of the panelists was thinking about how the performance levels were to be aligned across grade levels. Science and social studies panelists did not participate in this activity. Social studies and science panelists began by elaborating on the general PLDs, to better understand the PLDs that were currently written and how to define the differences in performance levels. Committees would use a flip chart to expand the definition of each performance level using the general PLDs. They started by first defining Level 2 (met), then Level 3 (basic), Level 1 (exceeded), and lastly Level 4 (apprentice). ELA and math committees participate in this activity as soon as they finished reviewing the alignment review. Note: ELA panelists had the extra task of writing SS PLDs for both reading and writing. The SS PLDs for science were already written, which was not known until the beginning of the meetings, so the science committees reviewed the currently written SS PLDs using the same format of the script as the other subjects.

The panelists then reviewed the NAEP subject and grade-level specific PLDs to better understand how to frame their writing of the subject-specific PLDs. Committees then began writing subject and grade-level specific PLDs starting with the Level 2 (met) category. They would first start with the lower of the two grades they were reviewing, and then write the PLDs for the upper grade. This same process was followed for Level 3 (basic), Level 1 (exceeded), and lastly Level 4 (apprentice). One additional task was that once a new level of PLDs was written, the panelists were asked to compare and contrast the PLDs from the level above and below (e.g. comparing met to basic). This was to help them see how well they were aligning their PLDs across grades as well as within a performance level. Once the committees wrote all four levels, they were asked to review all four levels using the pairs of grades in their committees. They were allowed to make any edits at that time. At the end of this task, the social studies and science committees were finished and completed their evaluations. Math and ELA teachers took a lunch break before coming back to two super-committees, one per subject, to do the vertical alignment across all six grades.

In the vertical alignment committees, there were 10 math panelists representing all six grades and seven ELA teachers representing all six grades. Panelists were asked to review the SS PLDs created by the other three committees. They were asked to think about what characterized students at each of the four performance levels, first within grade and then across grades. This was so the panelists could preserve the PLDs at each grade level, as well as to show how such performance was aligned across all six grades. The two committees struggled with this task, as each committee wrote their SS PLDs in different ways, even when given the NAEP PLDs as an example and a chart for filling in the SS PLDs within a grade and performance level. The mathematics committee appeared to come away with a relatively vertically aligned set of SS PLDs across grades, while ELA was not able to do so. There was an issue in ELA where one of the panelists was highly opinionated and would not let the committees move forward. Given the size of the panel was so small, this panelist's personality impacted the progress that could have been made in this committee.

The SS PLDs from the eight separate grade and subject-specific committees as well as the two sets of vertically aligned SS PLDs are forwarded to BAA and MEAP staff for review and edits, prior to being used for standard settings in January, 2006. Those documents are available at BAA. In addition, a complete report on SS PLDs is available at BAA.

## **6.2. Standard Setting**

The complete standard setting report is provided in *Appendix I*. This report includes information on the standard setting methods, panel selection, technical issues, results, and feedback, and details this process for mathematics, reading, writing, science and social studies. Since Fall 2008 administration, BAA adopted new performance level descriptor for MEAP. The performance levels are *Not Proficient*, *Partially Proficient*, *Proficient* and *Advanced*.

## **6.3. Revised Standards for Writing**

BAA administered revised Writing assessments for the first time in 2010, thus necessitating standard settings in order to determine students' proficiency and report results to the federal government. The outcome of the standard setting was the establishment of three cut scores that the BAA can use to classify students into one of four levels of performance: Level 4 = *Not Proficient*, Level 3 = *Partially Proficient*, Level 2 = *Proficient*, and Level 1 = *Advanced*.

One of the purposes of assessment is to establish clear guidelines for educational decision making. To this end, assessments require a process to establish standards that allow teachers, administrators, policy makers, and parents to make statements about the level of proficiency of individual students and groups of students. This process typically amounts to making determinations as to what levels of student performance are judged to be sufficient to meet performance standards on the test. Providing the context for this process are the performance level definitions (PLDs) that have been established by the BAA. This critical aspect of standard setting must be addressed in advance of the meeting in order to ensure that the resultant cut scores reflect established and agreed upon descriptive criteria of student performance.

The standard-setting process relies on expert judgments, much like the process of assigning grades to student work. In order to adequately consider the Writing standards and PLDs, it is critical that the standard-setting panelists have the necessary content area knowledge and student understanding to make informed judgments about student performance. These panelists must also represent the more general body of Michigan educators. Special care must be taken to ensure adequate professional, gender, racial/ethnic, and geographical representation on the standard-setting panels.

When deciding upon an appropriate standard-setting method it is essential to consider the composition of the assessment and how it is to be scored. The following outline provides that information.

The total number of items is 25, comprised of:

- 16 multiple-choice (MC)
- 8 analytic items (4 analytic scores given on each of two tasks, informational and narrative; see below)
- 1 holistically scored response whereby examinees rewrite and/or improve a given student writing sample

For each of the two constructed-response (CR) informational and narrative tasks, a student receives 4 analytic scores, and can earn a maximum of 15 points on each of these two tasks. The items (i.e., rubric labels) and possible scores for each are:

- Ideas 0-3, doubled, so 6 points possible
- Organization 0-3
- Style 0-3
- Conventions 0-3

In all, the total number of points possible for Writing is 50, made up of:

- 16 MC = 16 points
- 2 CR tasks, each with 4 analytic items of 15 points each = 30 points
- 1 holistically scored CR item = 4 points

### **6.3.1. Standard Setting Methodology**

Procedures that focus panelists on actual student work are gaining wide acceptance among statewide testing programs where increasing numbers of constructed-response (CR) or performance-type items are being used. The Body of Work (BoW) method is one commonly used approach which is appropriate in this context for several reasons. First, there is a mixture of both MC and CR items as well as a higher proportion of points coming from constructed response items. This mix lends itself well to illuminating the various ways in which a student can achieve a given sum score. With this method, the panelists' task is to classify student work into one of several performance categories defined to capture levels of performance as expressed by the PLDs. The method is holistic in that the panelists consider the whole of an individual student's constructed-response work and multiple-choice, i.e., all the items of a particular student for a grade. With a BoW sorting method, panelists review samples of student papers sampled to represent the full range of scores, and are asked, in essence, to sort these papers into four performance levels according to the quality of the students' work.

After initial consultation with BAA and the Technical Advisory Committee (TAC), Measurement Incorporated (MI) proposed a modified body of work procedure with three rounds of standard-setting. The procedure, with extensive input and advice from BAA, was a modified version of body of work in that panelists were given information in construct maps that showed the relationship of actual student work samples to a score scale that underlies the assessment. Prior to discussing how the implemented method differs in approach to the traditional BoW an explanation and an example of a construct map is provided to familiarize the reader with the concept.

A construct map is a tabular representation between the score scale and examinee and item data from the assessment. Construct maps are derived from item response theory (IRT) models, which are the scoring procedures used to scale all of the MEAP assessments. Versions of the general construct mapping framework that specifically emphasize particular components of construct maps have been used in previous standard-setting processes. These include item maps with Bookmark standard-setting, Reckase charts with Angoff standard-setting, and domain score charts with Mapmark standard-setting. Table 6.3.1.1 below shows an example of an expanded construct map.

Table 6.3.1.1: Example of a Construct Map (Mathematics)

Consequence Data (PAC)	Teacher's Students	Whole Booklets	Raters' Cut scores	Score Scale	Item Scores			Domain Scores		
					Item 1	...	Item 50	Number Sense	...	Algebra
...				...	..		...	...		...
14%		K, L	R1, R6	200	.91		.97	.95		.82
19%	Student A	M, N		197	.88		.96	.93		.81
24%	Student B and C	O, P	R2, R3	194	.83		.95	.91		.78
31%	Student D and E	Q, R	R4	191	.77		.94	.88		.74
36%	Student F, G and H	S, T		188	.70		.92	.85		.68
40%	Student I	U, V	R7	185	.63		.91	.82		.64
44%		W, X		182	.55		.89	.79		.59
48%	Student J and K	Y, Z	R5, R8, R9	179	.48		.86	.73		.55
53%		AA, BB		176	.42		.83	.66		.49
59%	Student L	CC, DD	R10	173	.37		.79	.65		.48
...				...	...		...	...		...

**Note:** The quantities in this table were contrived based on information that would be available when applying IRT. The letters in the Whole Booklets column correspond to booklets or score profiles. The letters under the Teacher's Students column represent students in the teacher's classroom. The abbreviations in the Raters' column show the location of the standard-setting judges. The numbers in the item scores and domain scores column represent expected performance on items or in domains based on the IRT item or test characteristic curves.

Construct maps such as the one in Table 6.3.1.1 provide a clear indication of what it means to set a cut score at a specific level. For example, if the cut score is set at 185, such as where rater 7's cut score is located, then this level corresponds to 40 percent of the students being at or above the cut score, the performance of Student I, the whole booklets U and V, expected performance on item 1 of 0.63, expected performance on item 50 of 0.91, expected performance in algebra of 64 percent, and expected performance in number sense of 82 percent (Table 6.3.1.1).

In the traditional BoW Standard Setting Method, the focus of panelists is actually on the Whole Booklets column, although not presented in the form of the construct map and not explained as such, and its relationship to the score scale when providing cut score recommendations. Between rounds data are introduced that show the distribution of rater cut score recommendations from the previous round and each panelist's own cut score recommendations. As part of the last round of standard setting, panelists also have an opportunity to review the implications of their cut scores in the form of impact data. These data are based on the cumulative frequency distributions of student scores at each score scale value or above. Panelists use all of this information in deciding on their final cut score recommendations. The construct map was introduced into this process in a way to better organize and bring together all of the various data we expected panelists to use when making their recommendations.

As already alluded to, the goal of the standard-setting was to recommend performance thresholds or cut scores in the best interests of students and the overall educational process. These recommendations helped inform the Michigan State Board of Education as it established performance standards for the statewide assessments.

### **6.3.2. Selection of Panelists**

Each person participating in the standard-setting process was selected for his or her qualifications as a judge of student performance based on various factors. Teachers, educators, community and business leaders, and subject area experts selected as panelists exemplified the required subject-area knowledge, teaching experience, and/or understanding of students necessary for an appropriate and comprehensive standard-setting study. Each panelist participating in the process represented the knowledge and understanding of his or her peers throughout the course of the process, lending a balance between diverse opinion and consensus.

To ensure balance and to appropriately represent a variety of opinions and positions, it was desirable to have a large and diverse group of panelists. The panelists were selected from an existing BAA database of available committee members, via solicitation of educational organizations and associations, and recommendations. A concerted effort was made to balance each panel on the basis of county representation, urban representation, and representation of schools serving various sizes of populations, gender, and race/ethnicity. The overarching goal of consensus in this forum was not the unanimous agreement of all parties, but the bringing together of individual divergent experiences to

form a common understanding of student performance that was truly larger, and broader, than its individual parts. It was desirable that the panelists selected for the standard-setting process represented the same diversity of peoples and demographics as the students assessed.

The number of panelists selected for each grade-specific committee was approximately 20-24, with small group exercises involving groups of 4 to 5. The panelists were, for the most part, educators familiar with the content area of Writing and the grade level. In addition to teachers, the educator groups included curriculum supervisors, principals, and district administrators. Furthermore, BAA solicited relevant community members and business leaders to serve as panelists. In order to achieve the requisite 20 panelists for each grade standard-setting panel, approximately 25 individuals per grade were invited to participate in the process to allow for the possibility that some people were not available during the designated time period.

### **6.3.3 Standard Setting**

The standard setting took place January 18-January 20, 2011. Although there are numerous activities planned, the three-day format provided the panelists ample time to adequately carry out the process as intended as well as enough time to deliberate within and between grade level bands. This latter step is important in ensuring that vertical articulation in the performance standards mirrors the vertical articulation inherent in the content standards driving the PLDs.

The training on January 18 included a thorough introduction to standard setting and the performance level descriptors (PLDs) for Not Proficient, Partially Proficient, Proficient, and Advanced. The final activity of the day was a presentation of a modified body of work (BoW) procedure that panelists would use. In this procedure, panelists evaluate student work samples that have been assembled into a set ranging from low to high total score. The panelist's task is to review work samples one at a time until finding the work sample that would just barely qualify as Partially Proficient, then identifying the work sample that just barely qualify as Proficient, and finally, the work sample that would just barely qualify as Advanced. The panelist then notes the score for each of the three identified work samples on a form. The cut score for a given level is taken as the median score for that level across all panelists assigned to that grade.

#### **6.3.3.1 Round 1**

Panelists were given a set of approximately 50 scored and ordered student papers and a construct map that helped them sort the papers and decide on their cut score recommendations. Each paper contained all responses to the assessment from a particular student. The papers were selected from a representative sample of the state. The papers spanned the range of the total test scores that students can receive and covered a distribution of scores for the CR items. Panelists provided their cut scores recommendations by separating the papers into groups that represented *Not Proficient*, *Partially Proficient*, *Proficient*, or *Advanced* student performance by examining the student work samples and the construct map.

Panelists started with the paper with the lowest score and asked themselves if they thought that student work sample was indicative of just barely Partially Proficient performance. If they believed that it was not indicative of just barely Partially Proficient performance then they moved to the next paper. As they moved from one paper to the next they were instructed to examine the construct map to see the score that the paper received. It was made clear to the panelists that moving from one paper to the next did not always represent the same change in achievement. In fact, in some cases moving from one paper to the next may have resulted in no change in achievement at all because the papers received the same score. When panelists reached a point where they thought they had found a paper that represented just barely Partially Proficient performance, they were instructed to examine that paper closely as well as a couple of papers above and below it in the construct map. From these papers and the scores that they received on the construct map, panelists were instructed to find the location on the score scale in the construct map that they thought best separated Not Proficient from Partially Proficient performance. The cut score that they recommended would have been a score from the score scale in the construct map. The cut score could have been at the location of a particular paper or set of papers or it could have been between several papers. They recorded this cut score in the construct map and on their separate rating sheet. It was made clear to panelists that the cut score they recommended was their best judgment of a cut score that separated examples of what they considered to be Not Proficient performance from Partially Proficient performance. Panelists were instructed to draw a line in the construct map where they placed their cut score and write the words Not Proficient below the line and Partially Proficient above the line as well as writing the word “Cut Score” in the Not Proficient/Partially Proficient column. This should have helped solidify for panelists how the different categories of performance were represented.

The panelists then proceeded to examine the test booklets directly above the cut score that they recommended to separate Not Proficient and Partially Proficient performance. Their attention turned to asking themselves whether each student work sample represented just barely Proficient performance. The process that they used to recommend this cut score was similar to the process that they used to identify the cut score to separate Not Proficient and Partially Proficient performance. Again, they were instructed to examine the paper and the score that the paper received in the construct map as they moved from paper to paper. When they found a paper that they thought was indicative of just barely Proficient performance they were instructed to examine a couple of papers directly above and below this paper in the construct map to come up with their best judgment for a cut score. It was made clear to the panelists that the cut score they recommended was their best judgment of a cut score that separated examples of what they considered to be Partially Proficient performance from Proficient performance. Panelists were instructed to draw a line in the construct map where they placed their cut score and to write the words Partially Proficient below the line and Proficient above the line as well as writing the word “Cut Score” in the Partially Proficient/Proficient column.

The process panelists used to identify the Proficient/Advanced Cut score was exactly the same as the process that they used to identify the other two cut scores.

The test booklets in whole booklets column are the student work samples that the panelist is reviewing as part of Round 1. The column labeled Not Proficient/Partially Proficient Cut Round 1 is where they recorded this cut score. The column labeled Partially Proficient/Proficient Cut Round 1 is where they recorded this cut score. The column labeled Proficient/Advanced Cut Round 1 is where they recorded this cut score. The panelists were also asked to rewrite their recommended scale score cuts for the three cut score placements on the rating form

It was pointed out to the panelists that the placement of the cut scores should correspond to the location on the score scale that they think best separates different categories of performance. It was possible for the panelist to select cut scores that were not at the location of one of the booklets as well as at the location of a particular test booklet. The reason for allowing the panelist to provide such cut score recommendations was to allow the panelist to make recommendations of locations where there may be score gaps (i.e., locations where there are not any samples of test performance). This minimized the impact of score gaps on the cut scores recommendations and potential cut score bias in the procedure. In addition, panelists often want to set their cut scores in between two papers and this allowed them to do that. The cut scores for the group of panelists in this round were the median of the panelists' cut score recommendations in each of the cut score recommendation columns.

The construct maps and rating sheets were collected and the cut scores were tabulated. The results from Round 1 were presented to the panelists using a construct map that showed the distribution of cut score recommendations and their own judgments. Because, each panelist was assigned a specific rater number, it was clear to panelists what rater number they were when the feedback was provided. They did not know the rater numbers of the other panelists.

Panelists were divided into groups of approximately 4-5 members in order to participate in group discussions of their ratings and the cut scores. Panelists were allowed to offer explanations for their cut scores and discussed their conclusions. No group consensus was pursued. These discussions were focused on the student work and performance description definitions. This step was used only to inform panelists of fellow panelists' rationales used to separate the work samples into different categories of performance and in how they arrived at their cut score judgments. Afterwards, panelists were asked in written form if they felt they understood the process and felt comfortable proceeding with the next round of standard setting. If panelists needed further review of procedures and/or definitions, those were provided. Panelists were asked to fill out a readiness and feedback form on their understanding of different components of the standard setting process from Round 1.

### 6.3.3.2 Round 2

The student papers used in Round 2 were more targeted than those in the first round. MI staff selected papers based on the cut scores that emerged from the ratings in Round 1. Panelists also received new construct maps with updated information and information on the new papers that they were asked to classify and used to make their new cut score recommendations.

Panelists were given sets of 30 student papers to use to provide their standard setting judgments (2 papers for each of 5 cut points surrounding the 3 cut scores). Each student paper contained all responses to the assessment from a particular student. These student work samples were different papers from those the panelists received in Round 1. The sample papers for this round of rating were selected to target a range of test scores that are near the cut scores which were indicated in Round 1.

Approximately 10 papers targeted the borderline between *Advanced* and *Proficient*, 10 papers targeted the borderline between *Proficient* and *Partially Proficient*, and 10 papers targeted the borderline between *Partially Proficient* and *Not Proficient*. The sets of targeted papers were selected by providing 2 additional papers at each of the recommended cut scores (or as close to them as possible) from Round 1 and 2 additional papers for each of the next two available scale score locations above and below each of the recommended cut scores (or as close to it as possible).

The process that panelists used to determine the cut scores in Round 2 were similar to the process that they used in Round 1 only the numbers of papers that they examined were less and more targeted. Panelists started with the paper with the lowest score surrounding the Not Proficient/Partially Proficient cut score and asked if the student work sample was indicative of just barely Partially Proficient performance. If they believed that it was not indicative of just barely Proficient performance then they moved to the next paper. As they moved from one paper to the next they again examined the construct map to see the score that the paper received. It was made clear to the panelists that moving from one paper to the next does not always represent the same change in achievement. It was also made clear that these papers are different than the papers that they rated in Round 1, but many of the papers represented similar types of performance. When panelists reached a point where they thought they had found a paper that represented just barely Partially Proficient performance, they were instructed to examine that paper closely and a couple of papers above and below it in the construct map. From these papers and the scores that they received on the construct map as well as the relationship of these papers to the sample of performance from Round 1, panelists were instructed to find the location on the score scale in the construct map that they thought best separated Not Proficient from Partially Proficient performance. The cut score that they recommended again had to be a score from the score scale in the construct map. The cut score they recommended again could have been at the location of a particular paper or set of papers or it could have been between several papers. It could have been the same cut score as the previous round or it could have been higher or lower than their previous cut score. They recorded this cut score in the construct map and on their separate rating sheet. It was made clear to

panelists that the cut score they recommended is their best judgment of a cut score that separated examples of what they considered to be Not Proficient performance from Partially Proficient performance. Panelists were instructed to draw a line in the construct map of where they placed their cut score and write the words Not Proficient below the line and Partially Proficient above the line as well as writing the word “Cut Score” in the Not Proficient/Partially Proficient column.

This process was repeated for the Partially Proficient/Proficient cut score and the Proficient/Advanced cut score.

At the end of the process, panelists had examined all of the student work samples from Round 2 and had written the words “cut score” to signify their Round 2 cut score placements in the construct map. These cut score recommendations were also collected on the separate rating so that the cut scores from this round could be determined.

The construct maps and rating forms were collected and the cut scores were tabulated using the median of the individual panelists’ cut score recommendations. The results from Round 2 were presented to the panelists using a construct map that showed the distribution of cut score recommendations and their own judgments. Again, because each panelist was assigned a specific rater number, it was clear to panelists what rater number they were when the feedback was provided. They did not know the rater numbers of the other panelists.

Panelists were allowed to offer explanations of their classifications and cut scores and discussed their conclusions. No group consensus was pursued. These discussions were focused on the student work and performance description definitions. This step was used primarily to inform panelists of fellow panelists’ rationale in making classifications and in how they arrived at their cut score judgments.

Panelists were asked in written form if they felt they understood the process and felt comfortable proceeding with the next round of standard setting. If panelists needed further review of procedures and/or definitions, those were provided. Panelists were asked to fill out a readiness and feedback form on their understanding of different components of the standard setting process from Round 2. Following Round 2 MI staff calculated impact data and provided this additional information to panelists.

### **6.3.3.3 Round 3**

Panelists received reports summarizing their individual ratings and the group cut scores after Step 6. They also were provided with statewide performance i.e., data to judge the impact of group cut scores. These data were based on virtually the entire population of Michigan’s grades 4 and 7 students tested. This information was again reflected in a construct map. The point of putting the information into the construct map was that it made it explicit to panelists how the student work samples, the score scale, and the frequency distribution of statewide performance were related.

Panelists were given data that gave a perspective on the effect of their ratings. They received information on the percentage students that were at or above the cut score for each of the scale scores on the test. Instruction was given so panelists could properly interpret the impact data in light of the information that they had used in previous rounds to make their cut score recommendations. Panelists discussed the impact of the cut scores on the state. Panelists also discussed the proper use of the reference information.

It was made clear to the panelists how the percentage of students in each of the four performance levels could be determined from the percentage at or above cut score (PAC) column in the construct map. In particular, the percentage of students in the Advanced category was the percentage in the PAC that corresponded to the Proficient/Advanced cut score location. The percentage of students in the Proficient category was the percentage of students at the location of the Partially Proficient/Proficient cut score minus the percentage of students at the Advanced category cut score. Similarly, the percentages of students in the Partially Proficient category was the difference between the percentage at the Not Proficient/Partially Proficient cut score location and the percentage in Proficient category. The percentage of students in the Not Proficient Category was found by subtracting 100 from the percentage at the Not Proficient/Partially Proficient cut score. The formulas for these simple computations are as follows:

Percent Advanced = Percentage at Proficient/Advanced Cut Score in PAC column  
Percent Proficient = Percentage at Partially Proficient/Proficient Cut Score in PAC column - Percentage at Proficient/Advanced Cut Score in PAC column  
Percent Partially Proficient = Percentage at Not Proficient/Partially Proficient Cut Score in PAC column - Percentage at Partially Proficient/Proficient Cut Score in PAC column  
Percent Not Proficient = 100 - Percentage at Not Proficient/Partially Proficient Cut Score in PAC column

Knowing these relationships allowed the panelists to see the impact of moving their cut score up and down on the score scale in the construct map as well as its relationship to the student work samples that they examined in rounds 1 and 2. The panelists were walked through how the percentages for the four performance categories were calculated based on the cut score recommendations determined from Round 2. This example allowed them to see how the process works and helped make it clear what moving the cut score did to the percentage of students that were in each of the four performance categories when they gave their final cut score recommendations.

In addition to the impact data, the respective grade groups also saw the cut scores from the other group. In other words, the grade 4 panelists saw the cut scores (from Round 2) generated by the grade 7 panelists and vice versa. Both groups also saw impact data from the MME Writing test. Seeing this information allowed the panelists from both groups to consider their respective scores, determine the degree of articulation, and consider this as they made their final recommendations in Round 3.

#### 6.3.3.4 Final Standard Determinations

Panelists recorded their cut score recommendations on their construct map and on the separate rating forms. They were allowed to make any changes they wished on the basis of the impact data and group discussions. Panelists were advised that this was the last round of adjustments. At this point, panelists were not reviewing student work samples. They were reviewing all of the information from the previous rounds of standard setting and the impact data to change the group cut scores as needed. It was made clear to panelists that their final cut score recommendations should represent their best judgment of cut scores and be grounded in the PLDs that were developed to guide the standard setting process. The panelists specifically examined the PLDs and all of the information in the construct map when providing their final cut score recommendations.

The final cut scores were determined from taking the median of the panelist cut score recommendations from this last round of the process. Panelists were given evaluation forms to complete and return. Ratings of the process and open-ended comments were encouraged. Specific questions on the evaluation form asked about the standard-setting process, construct maps, the panelists' comfort with their cut score recommendations and other questions related to their experience.

After the standard setting, BAA reviewed the results with the Technical Advisory Committee (TAC) and presented the results to the State Board of Education. The Board members then made the final decision about the cut scores. The final cut scores are presented in Table 6.3.3.4.1 below.

**Table 6.3.3.4.1 Final Cut Scores for Writing**

Assessment	Grade	Partially Proficient	Proficient	Advanced
MEAP	7	666	700	733
MEAP	4	362	400	429

#### 6.4. Revised Proficiency Level Cut Scores

In 2011, a special study was conducted to identify new cut scores on the MEAP, where proficient is defined as being on track to succeed in a postsecondary educational experience. The Grade 8 MEAP scores were linked to The Grade 11 Michigan Merit Examination (MME) scores, which were also linked to freshman college grades to identify cut scores on the MEAP all grades. The work was accomplished by MDE and ACT, Inc. New cut scores were set in Mathematics, Reading, Science, and Social Studies. Writing was not included in the study because the MME writing cut score is already similar to the ACT writing college readiness benchmark.

The first set of cut scores was to set cut scores to represent being on track to succeed in a postsecondary educational experience (for MME) and being on track to success in the next grade level tested (for MEAP). The second set of cut scores was to represent being

advanced beyond being on track to succeed in the next level of education. The final set of cut scores was to represent a level of achievement below being on track to succeed in the next level of education.

Three types of links needed to be made in order to identify cut scores. The first is to link 11<sup>th</sup> grade MME scores to freshman college grades to identify cut scores on the MME. The second is to link MME scores to MEAP scores to identify cut scores on one or more grades of the MEAP. The third is to link MEAP scores in one grade to MEAP scores in another grade to identify cut scores on one the remaining grades of the MEAP.

This was accomplished by relating course grades from first-year college students enrolled in Michigan public postsecondary institutions (two- and four-year) to MME and MEAP scores.

All Michigan postsecondary institutions were asked to provide a list of first-year credit-bearing courses that they felt would be appropriate. The final list was reviewed and approved by MDE staff. Each course was assigned to a subject area (mathematics, reading, science, or social studies). Some courses were used for both reading and social studies. Using the final list, grades for courses were pulled by ACT for the Center for Educational Performance and Information (CEPI) grade file provided by MDE. The final file included 13 four-year and 26 two-year public institutions.

Students with first college enrollment dates of Fall 2009 and Fall 2010 were used in the study. These were the first cohorts that had both 11<sup>th</sup> grade MME scores and college grades. After matching and cleaning, the final sample size was 104, 691 records.

For each subject area, Signal Detection Theory (SDT) was used to generate a distribution of consistency classifications across MME test score by institution. The median consistency at each score was calculated across institutions and a logistic regression function was fit to this distribution to smooth the results. The MME scores with the highest median consistency were selected as the 11<sup>th</sup> grade MME college readiness cutoff scores. A score that gives the highest classification consistency also has a probability of success of 0.50 meaning that students with this score have a 50 % chance of receiving a B or higher course grade in the subject area. Partially proficient and advanced cutoff scores were selected as the scores at which students had a 33% and 67% chance of success. Appendix I provided more information about this special study on the new cut scores. New MEAP proficiency level cut scores for all subjects were described from Table 6.4.1. to Table 6.4.4

**Table 6.4.1 Final Cut Scores for Math**

<b>Assessment</b>	<b>Grade</b>	<b>Partially Proficient</b>	<b>Proficient</b>	<b>Advanced</b>
MEAP	8	809	830	865
MEAP	7	714	731	776
MEAP	6	614	629	675
MEAP	5	516	531	584
MEAP	4	423	434	470
MEAP	3	322	336	371

**Table 6.4.2 Final Cut Scores for Reading**

<b>Assessment</b>	<b>Grade</b>	<b>Partially Proficient</b>	<b>Proficient</b>	<b>Advanced</b>
MEAP	8	796	818	853
MEAP	7	698	721	760
MEAP	6	602	619	653
MEAP	5	501	521	565
MEAP	4	395	419	478
MEAP	3	301	324	364

**Table 6.4.3 Final Cut Scores for Science**

<b>Assessment</b>	<b>Grade</b>	<b>Partially Proficient</b>	<b>Proficient</b>	<b>Advanced</b>
MEAP	8	826	845	863
MEAP	5	526	553	567

**Table 6.4.4 Final Cut Scores for Social Studies**

<b>Assessment</b>	<b>Grade</b>	<b>Partially Proficient</b>	<b>Proficient</b>	<b>Advanced</b>
MEAP	9	899	928	960
MEAP	6	593	625	649

## CHAPTER 7: SCALING

### *Rationale Behind Test Scaling*

The basic score on any test is the raw score, which is the number of items correct. However, the raw score alone does not present a wide-ranging picture of test performance because it can be interpreted only in terms of a particular set of items. When new test forms are administered in subsequent administrations, other types of derived scores should be used to compensate for any differences in the difficulty of the items and to allow direct comparisons of student performance between administrations. Measurement Incorporated (MI), in cooperation with the Michigan Educational Assessment Program (MEAP), provides a derived scale score system for reporting performance on the MEAP.

### *Rasch Measurement Models*

The Rasch Partial Credit Model (RPCM) is used to derive the scale score system for the MEAP. The RPCM, an extension of the Rasch model, accommodates the constructed response tasks associated with the MEAP.

The advantage of using IRT models in scaling is that all of the items measuring performance in a particular content area can be placed on the same scale of difficulty. The further value of the one-parameter (Rasch) model over more complex IRT models is that the Rasch model assumes that for each raw score point there is only one ability. This relationship allows the Rasch difficulty values for the individual items to be used in computing a Rasch ability level for any raw score point on any test constructed from these items, thereby allowing for scaling and longitudinal comparability.

The Rasch Partial Credit Model (RPCM) is an extension of the Rasch model attributed to Georg Rasch (1960), as extended by Wright and Stone (1979), and Wright and Masters (1982). The RPCM is used because of its flexibility in accommodating multiple-response category data and its ability to maintain a one-to-one relationship between the derived (i.e., scale) and the underlying raw score scale. The RPCM is the underlying scale score system that will facilitate the equating of multiple test forms and allow for comparisons of student performance across years. Additionally, such an equitable scale will facilitate the transfer of equivalent performance standards across the years. The RPCM is defined via the following mathematical measurement model where, for a given item involving  $m$  score categories, the probability of person  $n$  scoring  $x$  on prompt  $i$  is given by:

$$P_{xni} = \frac{\exp \sum_{j=0}^x (B_n - D_{ij})}{\sum_{k=0}^{m_i} \exp \sum_{j=0}^k (B_n - D_{ij})},$$

where  $x = 0, 1, 2, \dots, m$ , and

$$\sum_{j=0}^0 (B_n - D_{ij}) = 0.$$

The RPCM provides the probability of a person scoring  $x$  on the  $m_i$  step of task  $i$  as a function of the person's ability ( $B_n$ ) and the step difficulties of the  $m$  steps in task  $i$  (see Masters, 1982, for an example).

## **7.1. Summary Statistics and Distributions from Application of Measurement Models**

Classical test theory (CTT) and item response theory (IRT) analyses are applied to MEAP data. Section 7.1.1 provides the summary of the CTT analyses, and section 7.1.2 is the summary of the IRT analyses.

### **7.1.1. Summary Classical Test Theory Analyses by Form**

The tables in *Appendix J* provide summary classical test theory analyses by form, grade and subject. The mean p-values for mathematics grade 3-8 are .67, .66, .61, .57 .58 and .51, and the mean point-biserial values range from .37 to .40 across grades. The mean p-values for reading are .73, .64, .69, .69, .69 and .70, and the mean point-biserial values range from .34 to .39 across grade. For science, grade 5 has mean p-values range from .57 to .60 whereas grade 8 has the mean p-values range from .49 to .54. The mean point-biserial values for science range from .30 to .32 and from .28 to .33 for grades 5 and 8, respectively. For social studies, the grade 6 and 9 have mean p values of .056 and .52 for grade 6 and 9 and the mean point-biserial values are .33 and .35, respectively. For writing, the mean p-values are .64 and .70 for grades 4 and 7, respectively and the mean point-biserial values are .44 for grade 4 and .41 for grade 7.

### **7.1.2. Item Response Theory Analyses by Form and for Overall Grade-Level Scales**

The assessments are analyzed using the Rasch Partial Credit Model and procedures implemented in WINSTEPS version 3.68.2. The statistical elements of the calibrating/scaling process are referred to as Rasch Calibration/Scaling as described in the WINSTEPS manual. Different versions of the same program may have subtle differences in how they implement the estimation routines, which on the surface appear to be consistent with past techniques. Because these changes may result in meaningful differences in estimation outcomes when applied to large testing populations, the version 3.68.2 was used for calibration activities.

#### **7.1.2.1. Step-by-Step Description of Procedures Used to Calibrate Student Responses, Using Item Response Theory**

The scaling design is referred to as a common item nonequivalent groups design (Kolen & Brennan, 2004). Bureau of Assessment and Accountability (BAA) build the test forms based on the test blueprint and available statistical information from data field tested in previous years. All assessments (mathematics, reading, science, social studies and writing) use an embedded matrix sampling design for building and replenishing the item

pool. For 2012, there are 5 forms for each subject. A sparse matrix that included all the scored items was created for each subject area by grade. A concurrent calibration was applied for each subject area by grade. Detailed procedures were provided in the sections below.

### **Calibration Steps for Mathematics, Reading, Science and Social Studies**

1. Run WINSTEPS without anchored items (free-run)
2. Do an anchor evaluation procedure:
  - a. Compute the ‘Mean-Mean’ *equating constant* =: *EQK* as the difference between the average of pre-equated Rasch values (scaled to 2010 operational year) and average of free-run Rasch values.
  - b. For all items, calculate the difference between the operational Rasch value + EQK and the pre-equated Rasch value.
  - c. Flag any item from consideration if the difference in absolute value is greater than **0.5**.
  - d. If there are no flagged items then stop, otherwise remove the flagged item with the maximum difference from the item list and *go to Step 2.a*.
3. The equated theta value is defined as:

$$\theta = \theta_{FreeRun} + EQK$$

In every step involved in the scaling and equating procedure, Assessment and Evaluation Services (AES) served as subcontractor for the Independent Psychometric Quality Assurance Review. AES reviewed and replicated all psychometric procedures connected to the scaling and equating of the assessments. The primary contractor, Measurement Incorporated, provided to AES all the same data which is provided to their psychometric unit. The prime contractor also provided AES with all necessary software settings, documentation, and the results of its own psychometric analyses for verification by AES. AES performed its analysis independent of the primary contractors work. The verification procedures are described in Appendix X.

#### **7.1.2.2. Summary Post-Calibration Score Distributions**

The tables in *Appendix K* provide summary post-calibration score distribution by form, grade and subject (scale score distribution and performance level percentage). The overall summary post-calibration score distribution by grade and subject is presented in Table 7.1.2.2.1.

Table 7.1.2.2.1  
Fall 2012 Administration –Scale Score Distribution by Grade and Subject  
With Performance Level Percentages

Subject	Grade	Form	Scale Score					% of Performance Level			
			N	Mean	SD	Min	Max	Not Proficient	Partially Proficient	Proficient	Advanced
MA	03	00	109640	332.03	22.52	209	415	35.68	23.39	36.90	4.03
MA	04	00	108147	431.51	25.72	283	537	38.41	15.41	37.90	8.27
MA	05	00	108066	530.09	32.39	363	660	37.08	17.13	40.18	5.62
MA	06	00	111602	625.1	28.7	471	758	39.08	20.73	34.71	5.49
MA	07	00	113859	725.09	29.82	572	862	37.62	24.11	32.26	6.02
MA	08	00	113382	820.06	29.55	681	950	38.94	26.33	26.55	8.18
RD	03	00	108915	333.32	26.9	189	418	11.59	21.83	57.37	9.21
RD	04	00	107250	433.52	28.17	283	536	7.05	24.77	63.15	5.03
RD	05	00	107425	536.54	29.21	388	630	11.55	17.97	57.87	12.61
RD	06	00	111096	631.04	26.28	490	724	14.34	17.3	45.56	22.8
RD	07	00	113675	728.67	30.54	576	826	15.77	22.04	48.03	14.16
RD	08	00	113253	828.33	25.17	689	916	9.68	24.44	53.42	12.47
SC	05	00	110759	523.8	25.94	356	668	52.13	34.71	8.25	4.91
SC	08	00	115748	821.1	23.22	670	966	57.91	26.16	11.93	4.01
SS	06	00	114647	611.54	21.51	481	727	20.46	49.76	26.12	3.65
SS	09	00	123315	913.8	24.86	779	1046	28.89	42.54	25.17	3.40
WR	04	00	107359	398.71	24.6	247	513	4.25	48.96	35.12	11.67
WR	07	00	113719	701.47	26.04	531	809	7.85	40.3	41.79	10.07

For mathematics, the mean scale scores for the overall grade level across grades are 332.03, 431.51, 530.09, 625.10, 725.09 and 820.06 with standard deviation between 22.52 and 32.39. Overall, the student performance is consistent across forms based on the observation from the scale score distribution and the performance level.

For reading, the mean scale scores for the overall grade level across grade are 333.32, 433.52, 536.54, 631.04, 728.67 and 828.33 with standard deviation between 25.17 and 30.54. Overall, the student performance is consistent across forms based on the observation from the scale score distribution and the performance level percentage.

For science, the mean scale scores for grade 5 and grade 8 are 523.80 and 821.10 with standard deviations of 25.94 and 23.22 for grade 5 and grade 8, respectively. Overall, the

student performance is consistent across form, based on the observation from the scale score distribution and the performance level percentage.

For social studies, the mean scale scores for grade 6 and grade 9 are 611.54 and 913.80, respectively. The standard deviations for grade 6 and grade 9 are 21.51 and 24.86, respectively. Overall, the student performance is consistent across form, based on the observation from the scale score distribution and the performance level percentage.

For writing, the mean scale scores for grade 4 and grade 7 are 398.71 and 701.47, respectively. The standard deviations for grade 4 and grade 7 are 24.60 and 26.04, respectively. Overall, the student performance is consistent across form, based on the observation from the scale score distribution and the performance level percentage.

Scale score histograms with the overlaid cut-score for the base forms by grade and subjects are displayed in Figures 7.1.2.2. Also the figures in *Appendix L* are the scale score histograms with the overlaid cut-score across form, grade, and subject. The visual displays share the same patterns as the tables in *Appendix K*; the student performance is consistent across form on scale score distribution and the performance level percentage.

Figure 7.1.2.2: Fall 2012 Administration Math 03 — Scale Score Distribution with Performance Level Outpoint Overlay

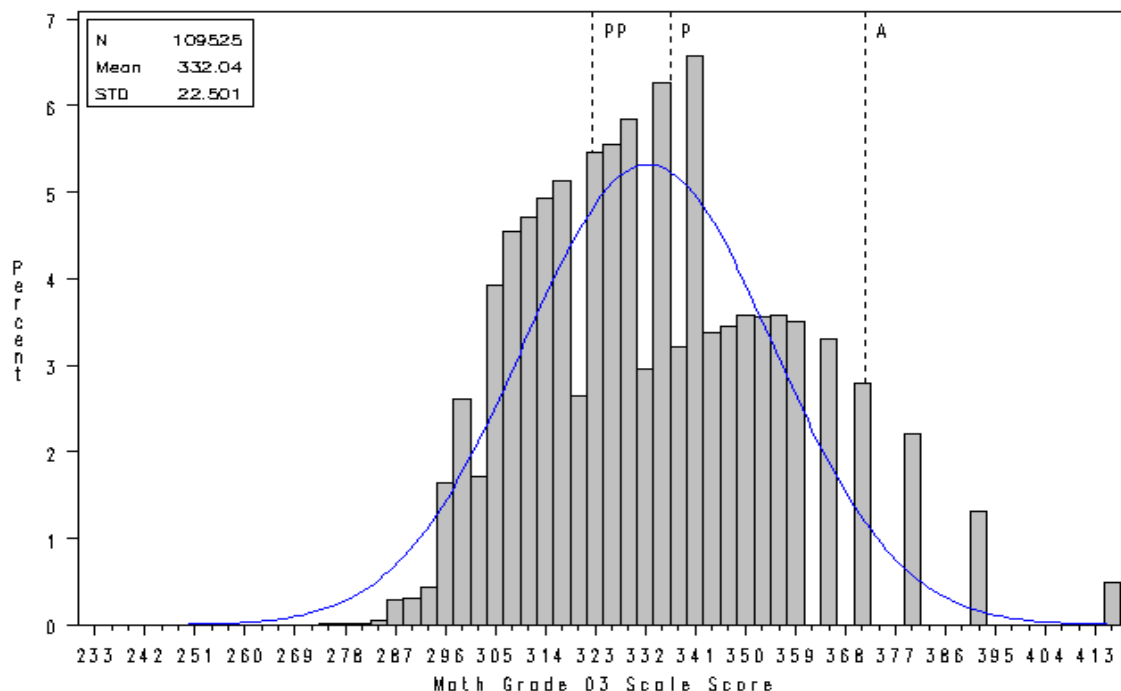


Figure 7.1.2.2: Fall 2012 Administration Math 04 — Scale Score Distribution with Performance Level Outpoint Overlay

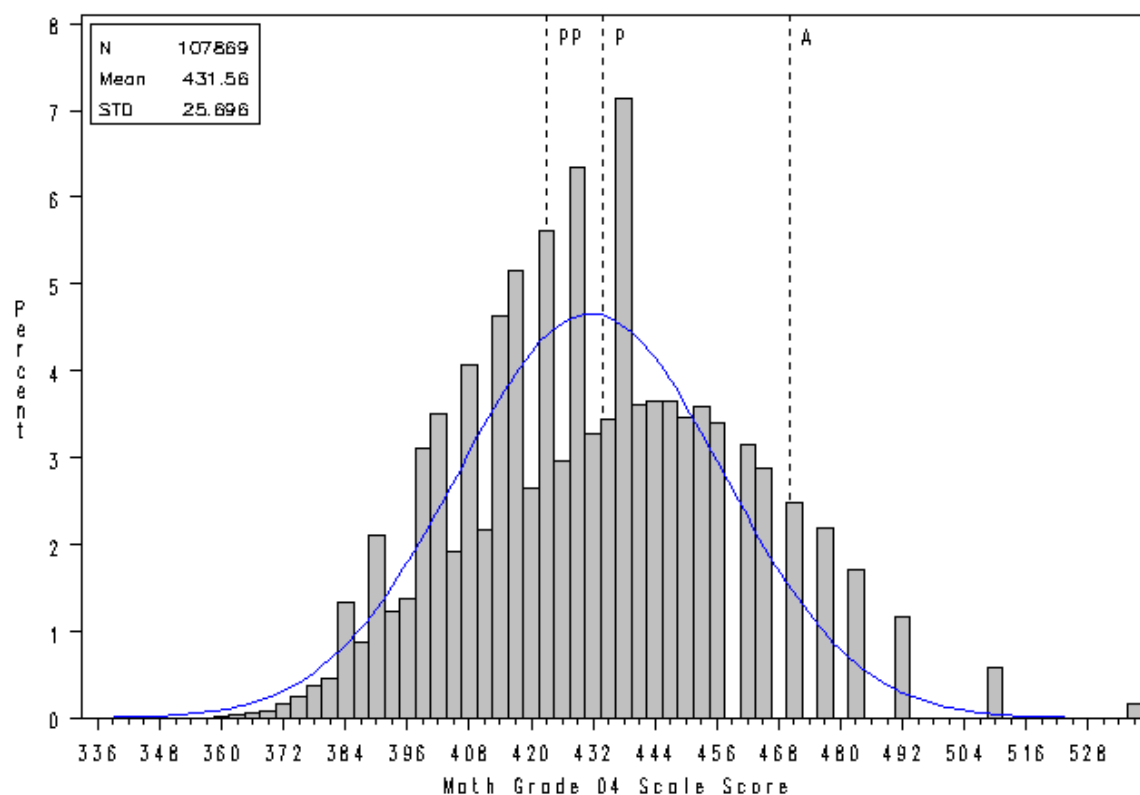


Figure 7.1.2.2: Fall 2012 Administration Math 05 — Scale Score Distribution with Performance Level Outpoint Overlay

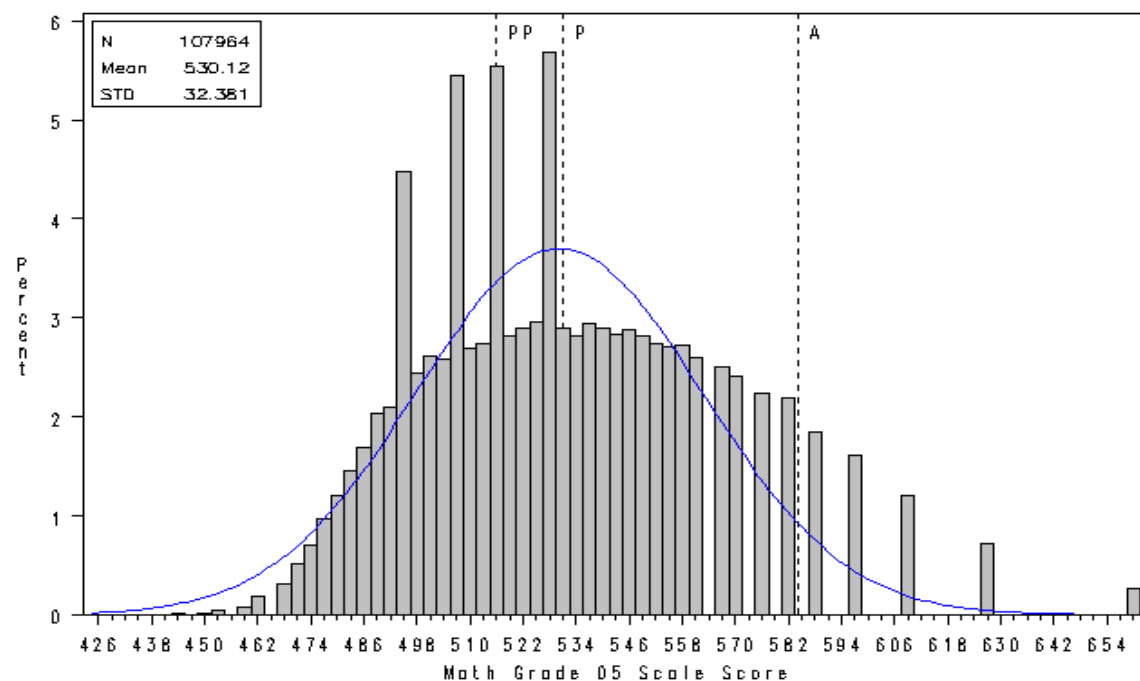


Figure 7.1.2.2: Fall 2012 Administration Math 06 — Scale Score Distribution with Performance Level Outpoint Overlay

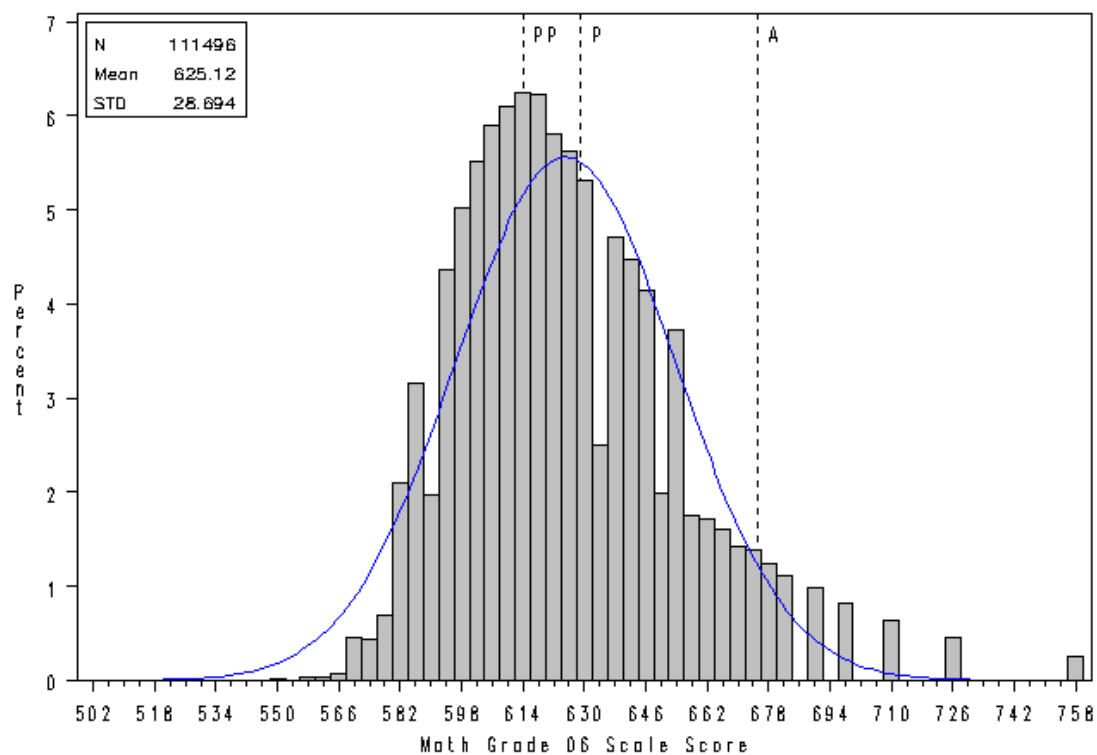


Figure 7.1.2.2: Fall 2012 Administration Math 07 — Scale Score Distribution with Performance Level Outpoint Overlay

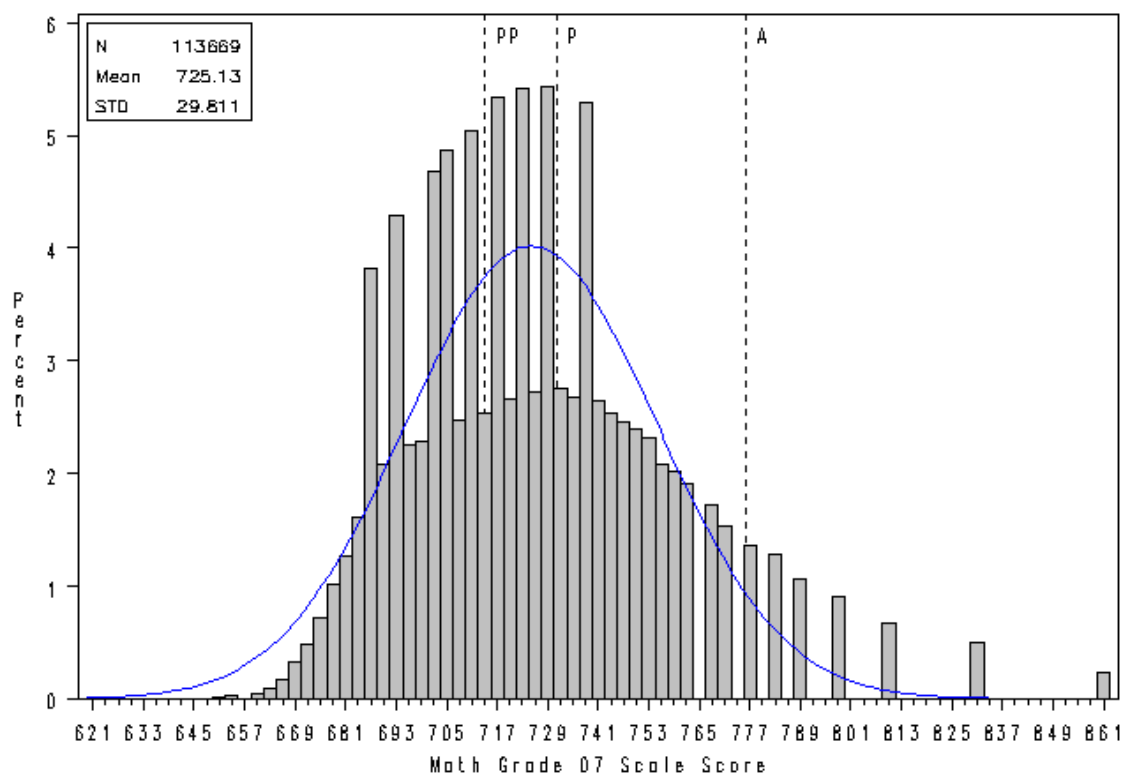


Figure 7.1.2.2: Fall 2012 Administration Math 08 — Scale Score Distribution with Performance Level Outpoint Overlay

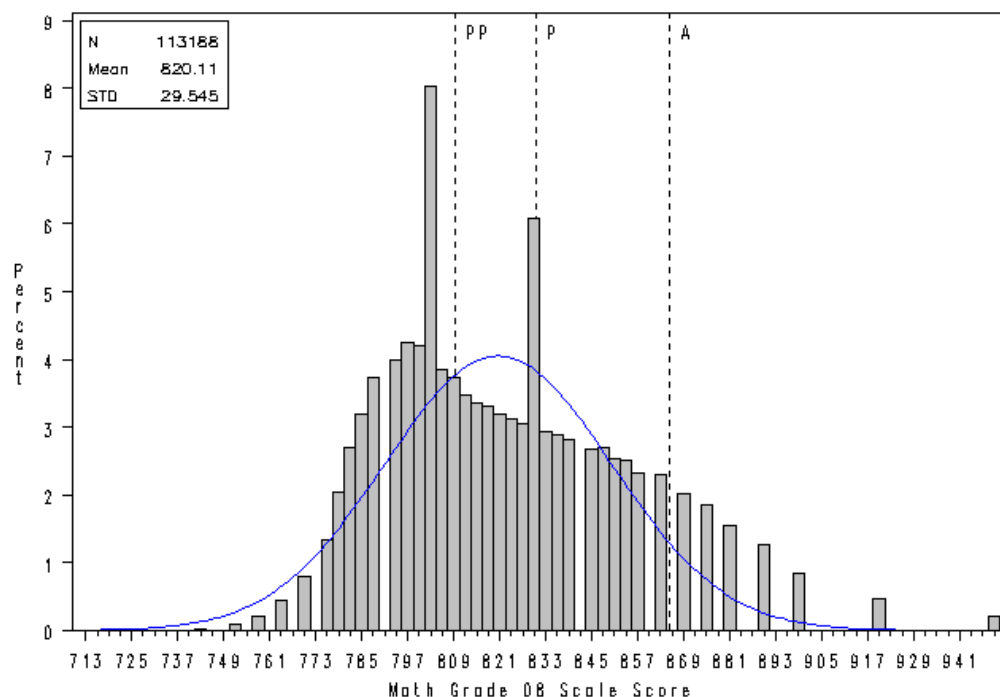


Figure 7.1.2.2: Fall 2012 Administration Reading 03 — Scale Score Distribution with Performance Level Cutpoint Overlay

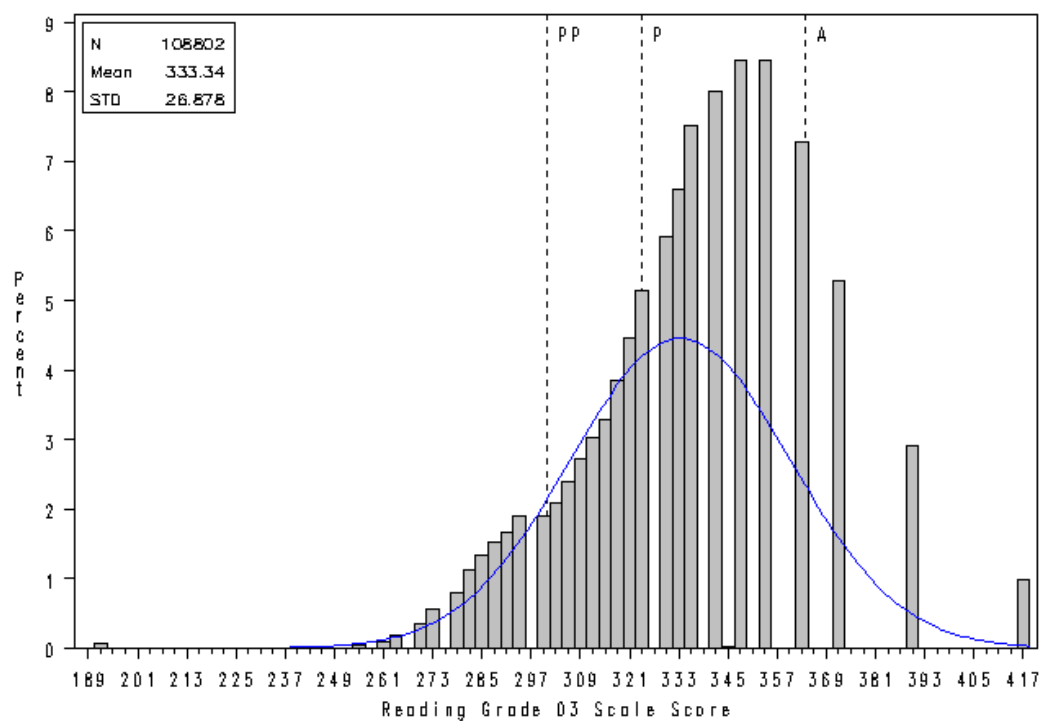


Figure 7.1.2.2: Fall 2012 Administration Reading 04 — Scale Score Distribution with Performance Level Cutpoint Overlay

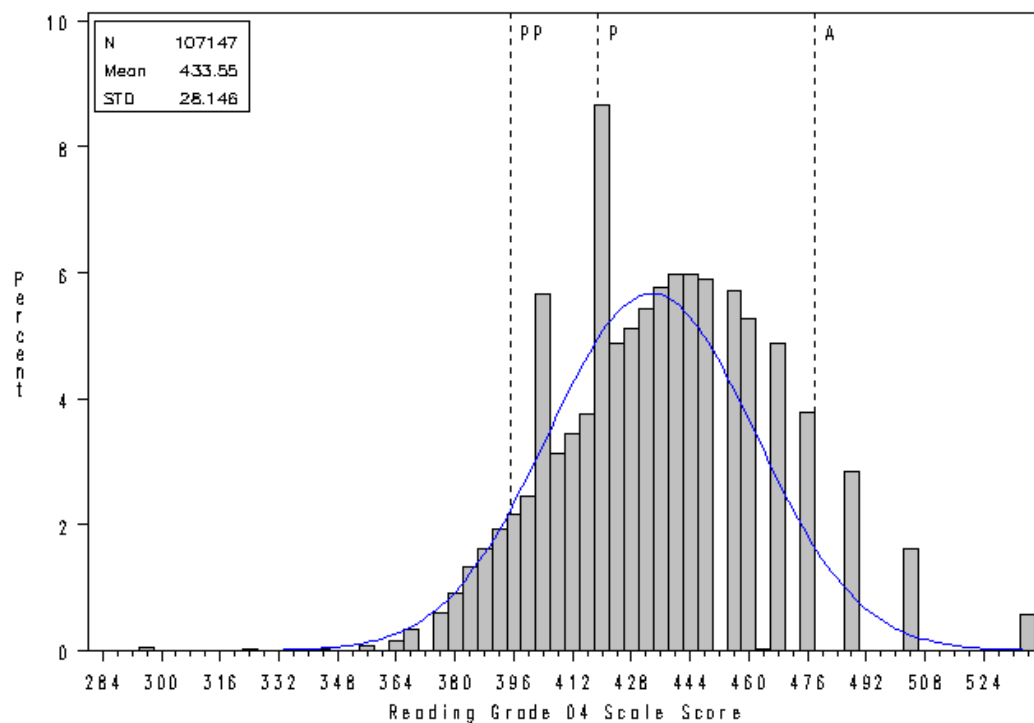


Figure 7.1.2.2: Fall 2012 Administration Reading 05 — Scale Score Distribution with Performance Level Cutpoint Overlay

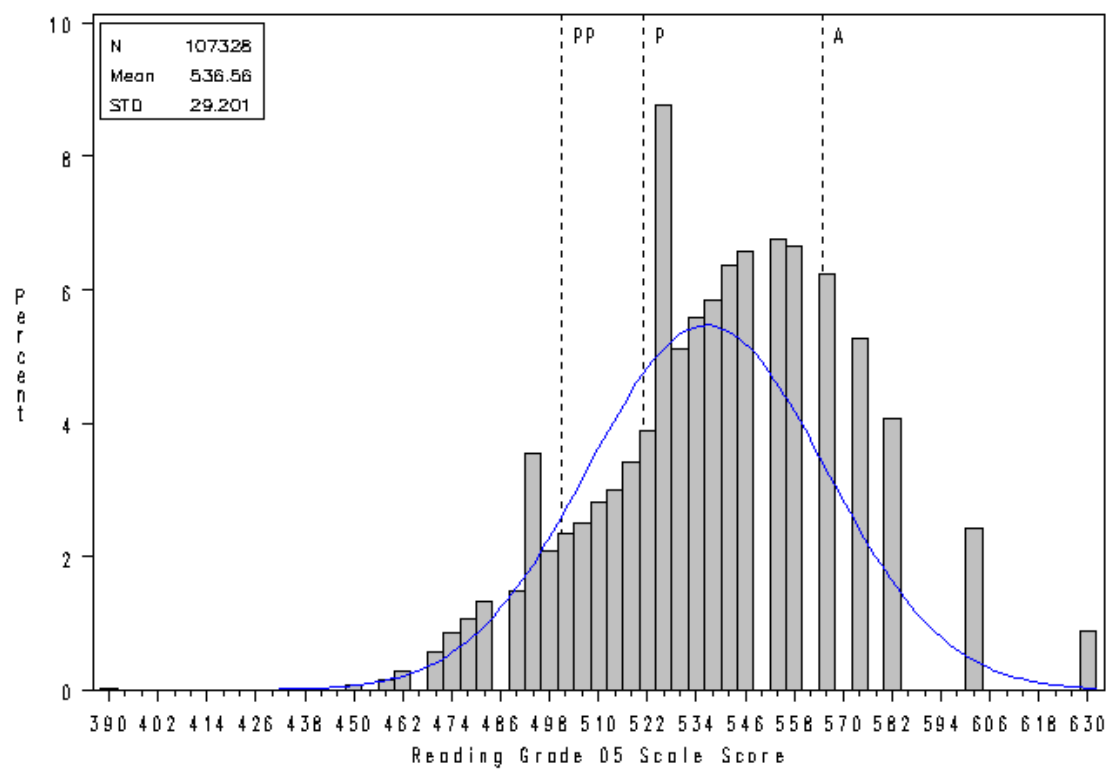


Figure 7.1.2.2: Fall 2012 Administration Reading 06 — Scale Score Distribution with Performance Level Cutpoint Overlay

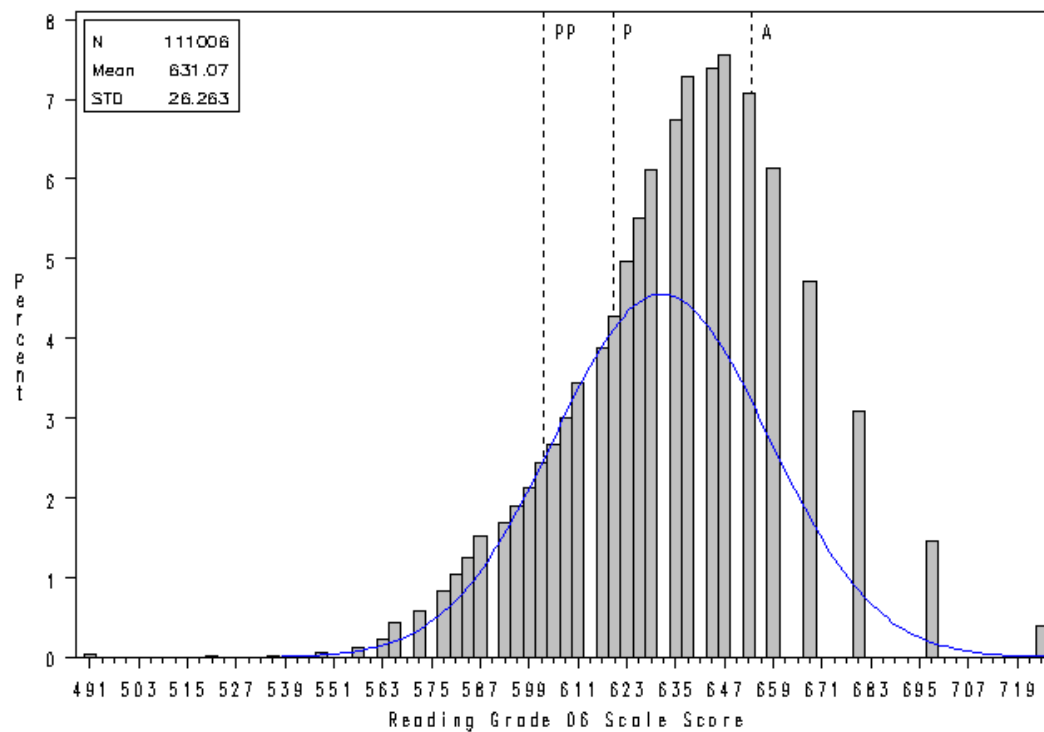


Figure 7.1.2.2: Fall 2012 Administration Reading 07 — Scale Score Distribution with Performance Level Cutpoint Overlay

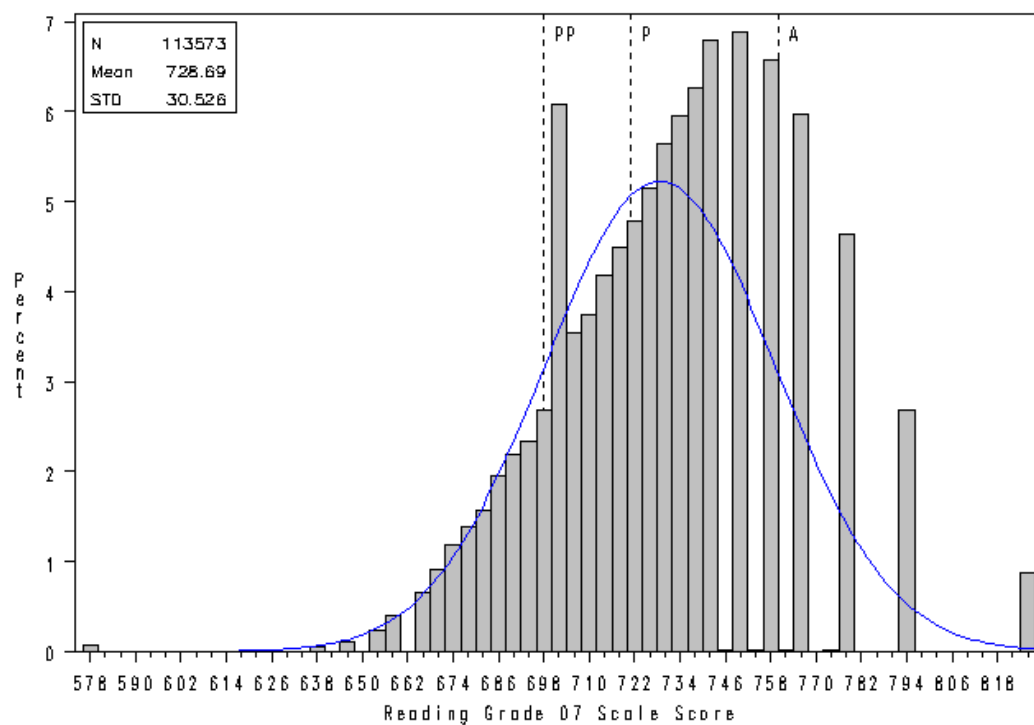


Figure 7.1.2.2: Fall 2012 Administration Reading 08 — Scale Score Distribution with Performance Level Outpoint Overlay

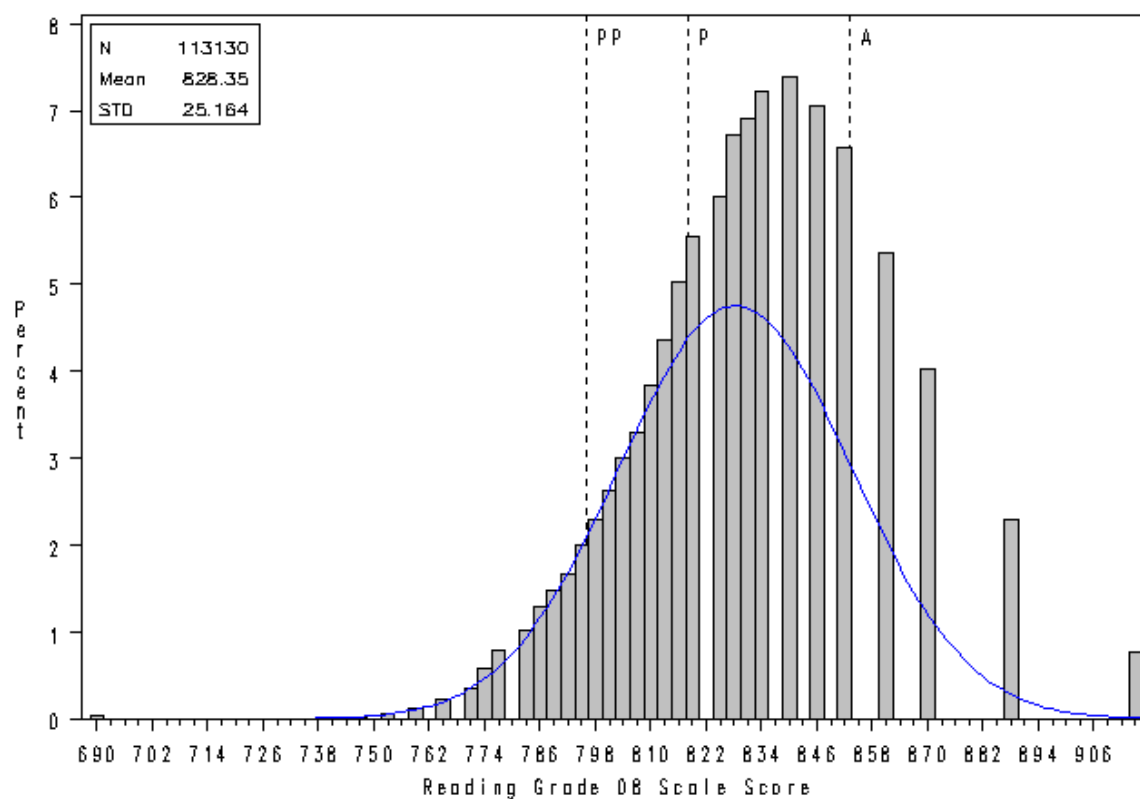


Figure 7.1.2.2: Fall 2012 Administration Science 05 — Scale Score Distribution with Performance Level Outpoint Overlay

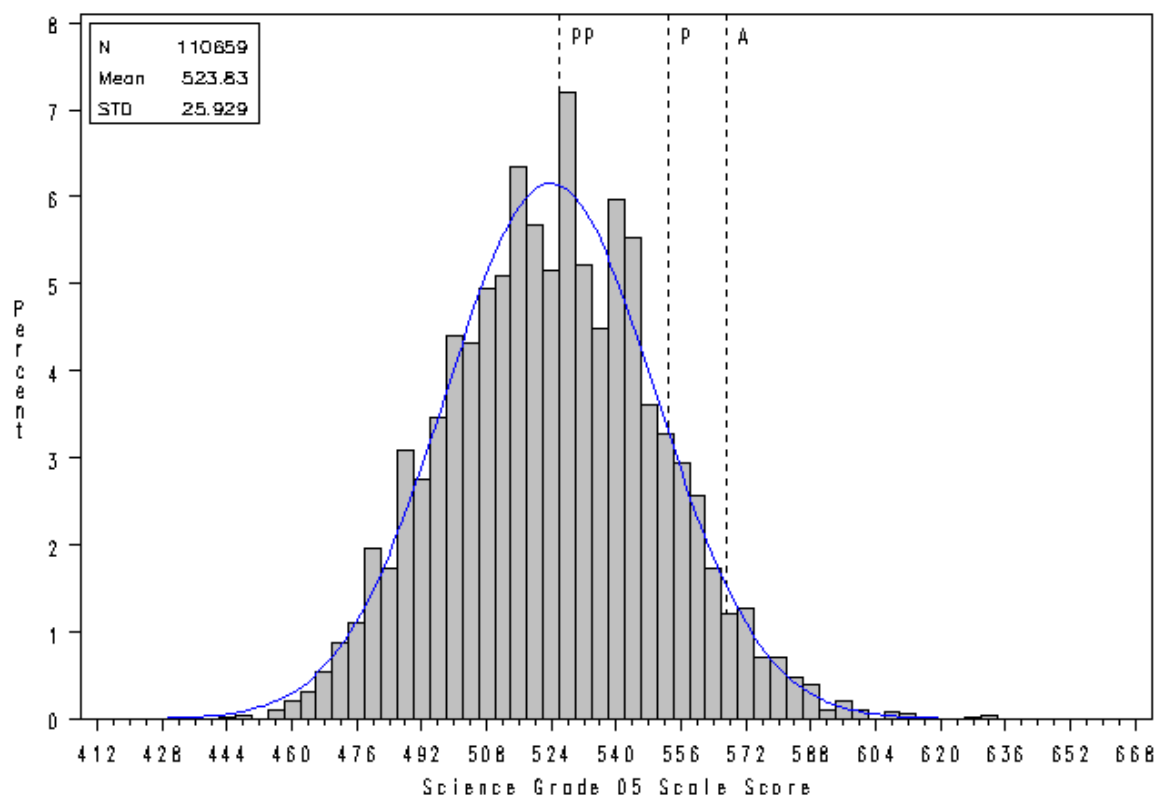


Figure 7.1.2.2: Fall 2012 Administration Science 08 — Scale Score Distribution with Performance Level Outpoint Overlay

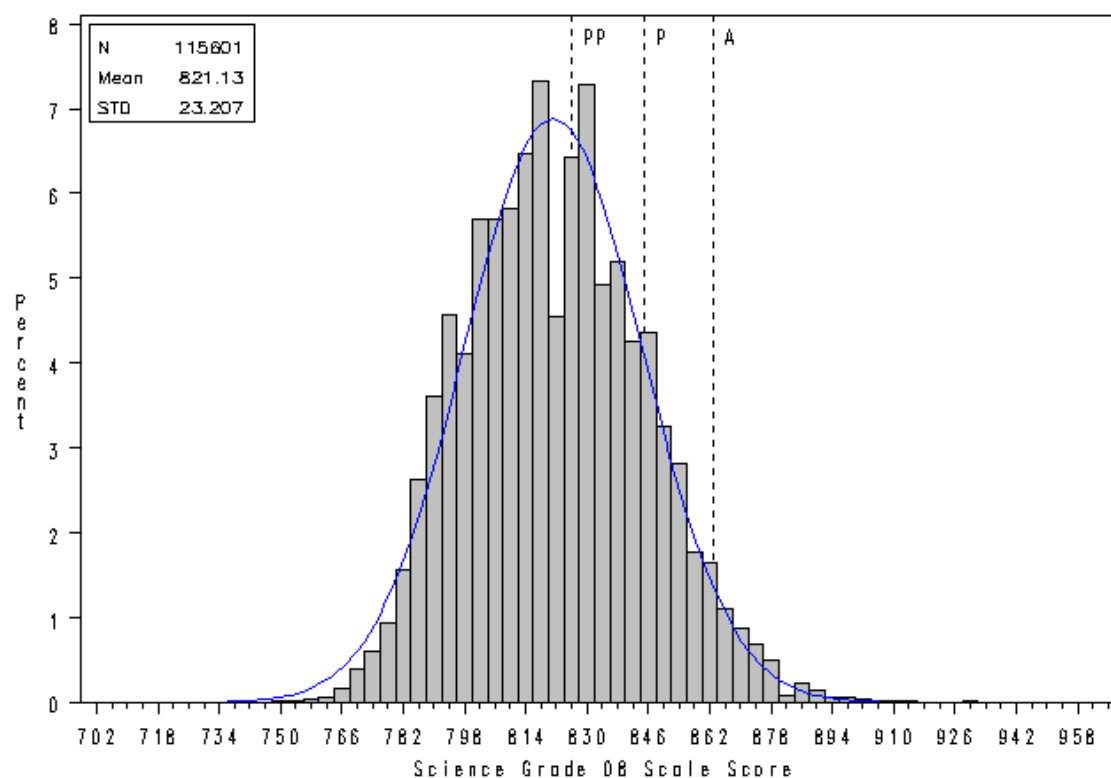


Figure 7.1.2.2: Fall 2012 Administration Social Studies 06 — Scale Score Distribution with Performance Level Outpoint Overlay

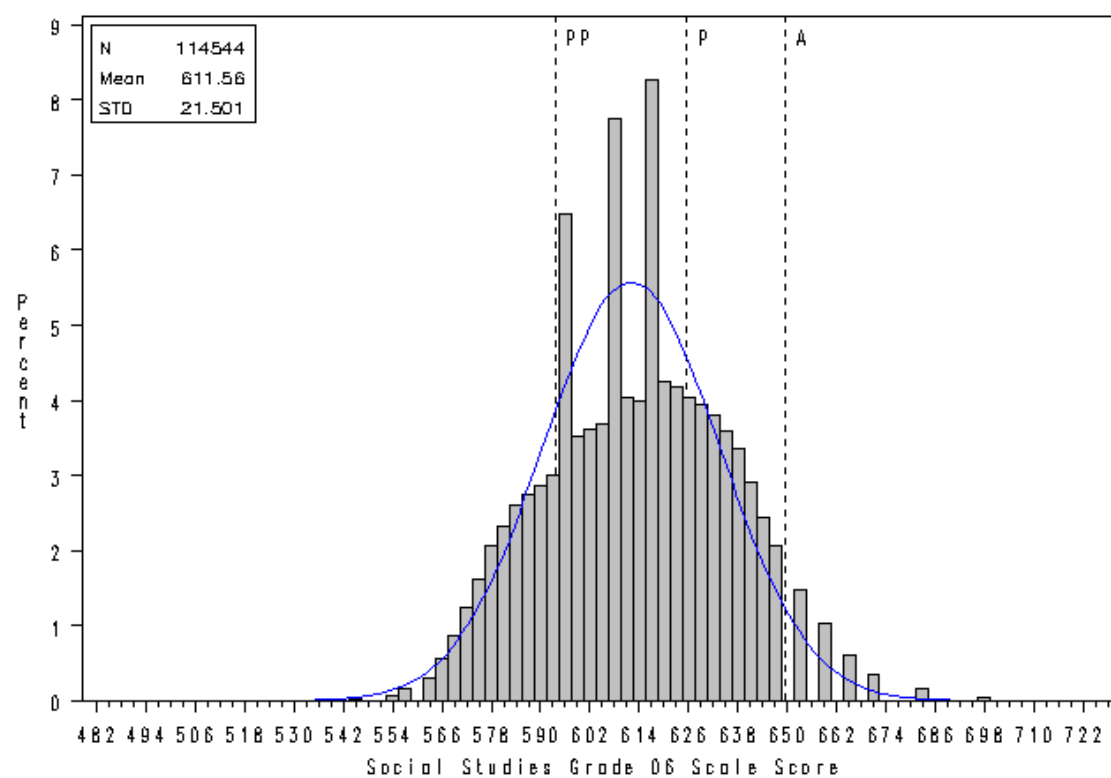


Figure 7.12.2: Fall 2012 Administration Social Studies 09 — Scale Score Distribution with Performance Level Outpoint Overlay

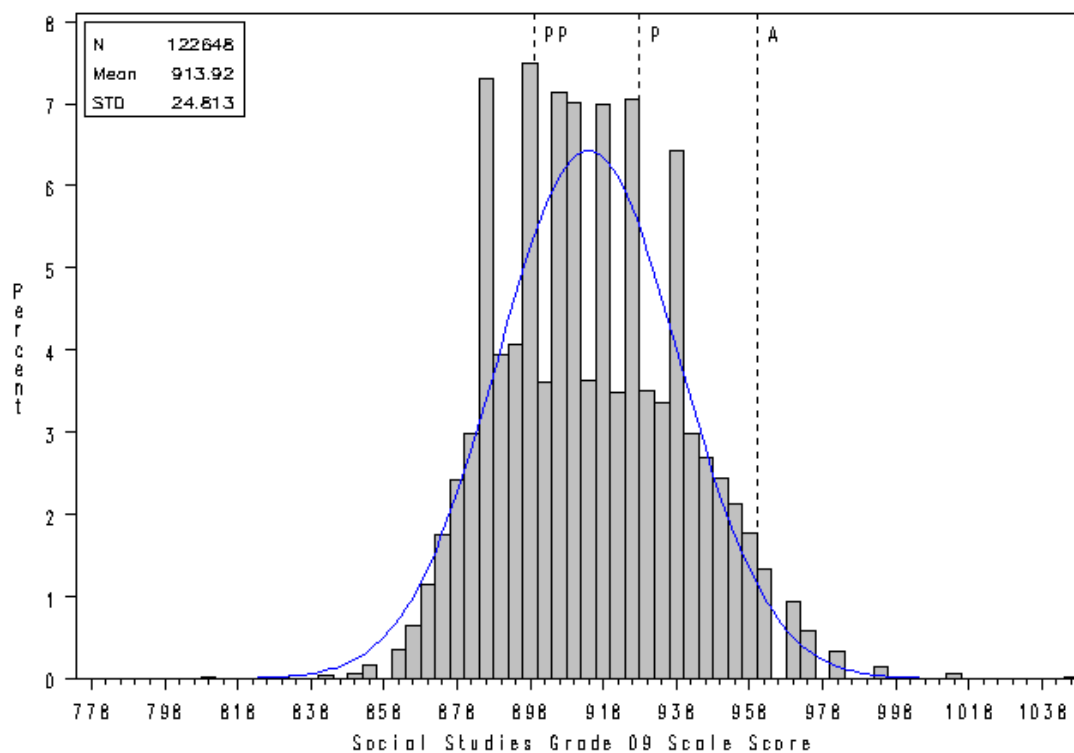


Figure 7.12.2: Fall 2012 Administration Writing 04 — Scale Score Distribution With Performance Level Outpoint Overlay

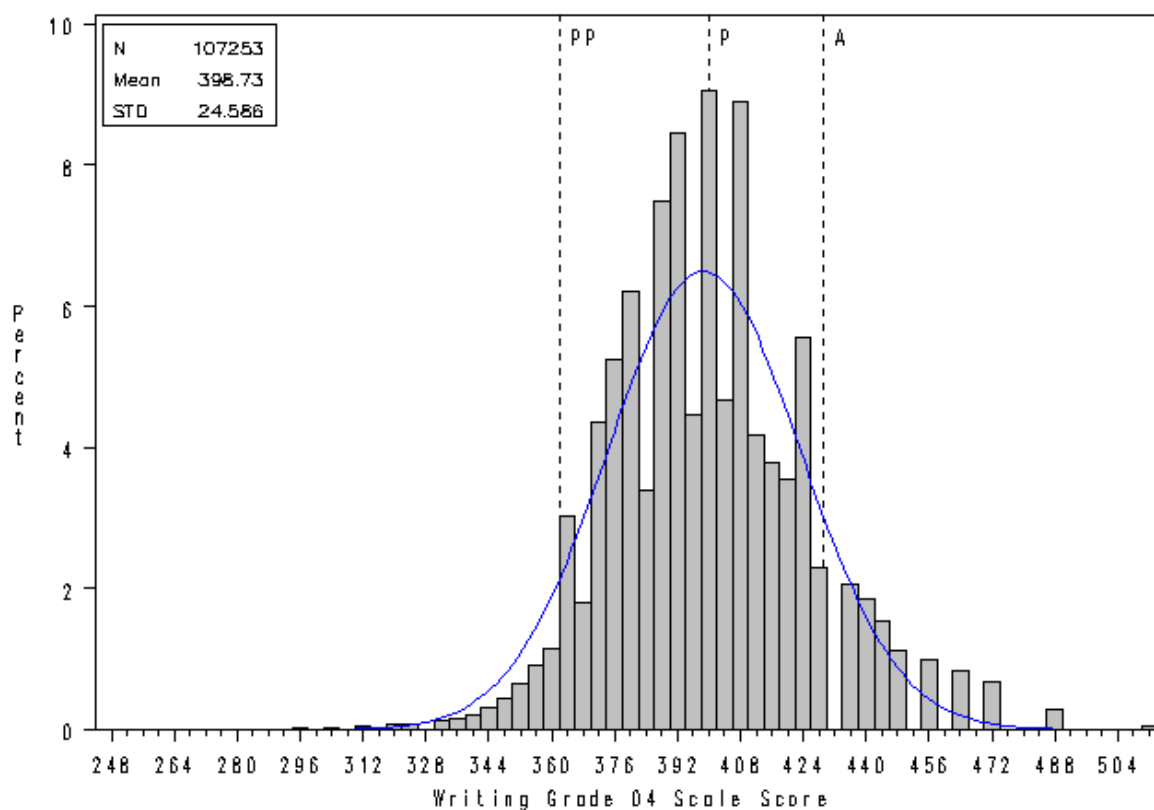
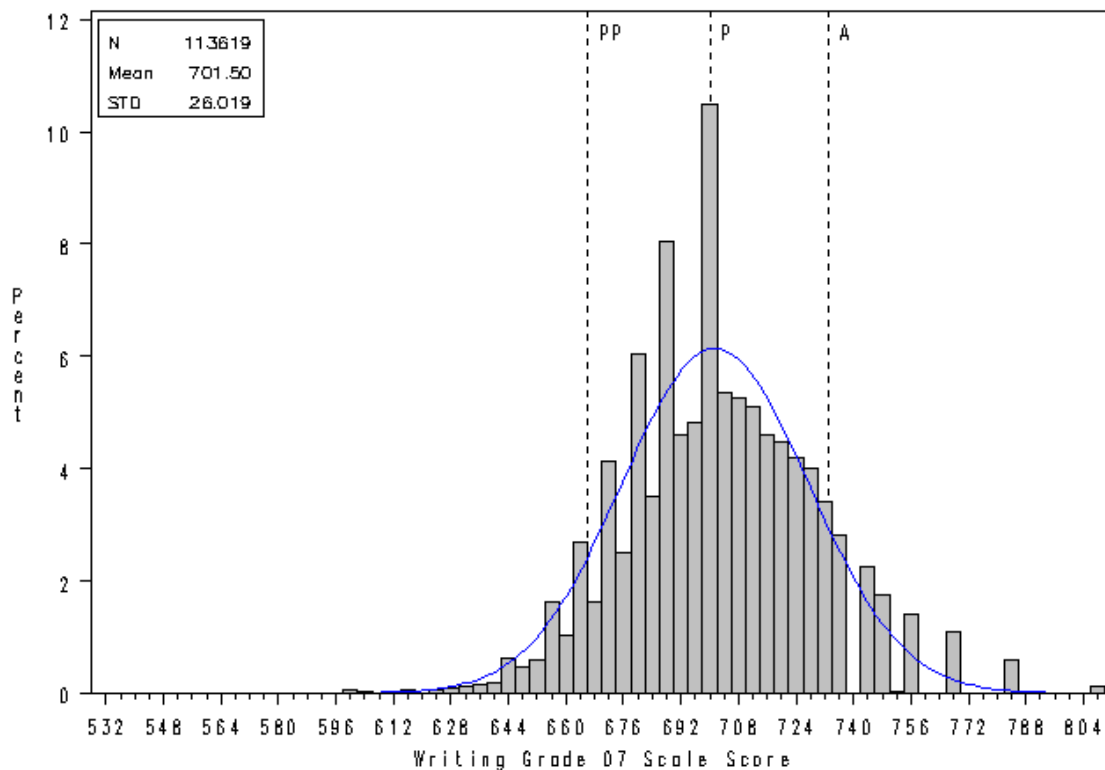


Figure 7.1.2.2: Fall 2012 Administration Writing 07 — Scale Score Distribution With Performance Level Cutpoint Overlay



### 7.1.2.3. Summary Statistics on Item Parameter Distributions and Fit Statistics

Item parameter distributions for grades and subjects are displayed in Figures 7.1.2.3. The tables in *Appendix M* provide summary statistics on item parameter distributions and fit statistics by form, grade, and subject.

For mathematics, the mean  $b$  parameter from grade 3 to grade 8 are 1.12, 0.93, 0.46, 0.73, 0.74 and 0.70 respectively, with standard deviation of 0.76, 1.00, 0.95, 0.91, 0.96 and 0.70. The mean infit statistics range from 0.98 to 1.00 with standard deviation between .10 and .14 across form and grade, and the mean outfit statistics range from 0.96 to 1.03 with standard deviation between .17 and .33.

For reading, the mean  $b$  parameter values from grade 3 to grade 8 are -0.18, 0.06, 0.04, 0.09, -0.09 and 0.39 respectively with standard deviation of 0.80, 0.74, 0.93, 0.83, 0.75 and 0.58. The mean infit statistics range from 0.99 to 1.02, with standard deviation between .09 and .10 across grades and the mean outfit statistics range from 0.93 to 1.01 with standard deviation between 0.16 and 0.20.

For science grade 5, the mean  $b$  parameter values range from 0.05 to 0.22 across forms with standard deviation between 0.86 and 0.97. The mean infit range from 0.99 to 1.01 and mean outfit range from 0.98 to 1.02. For science grade 8, the mean  $b$  parameter

values range from 0.28 to 0.53 across forms with standard deviation between 0.73 and 0.83. The mean infit range from 0.98 to 1.01 and mean outfit range from 0.98 to 1.03.

For social studies grade 6 and grade 9, the mean  $b$  parameter values are -0.08 and -0.05 with standard deviation of .62 and .58. The mean infit statistics are .99 and .98 for grades 6 and 9, respectively and the mean outfit statistics are 0.98 for both grades 6 and 9.

For writing grade 4 and grade 7, the mean  $b$  parameter values are -0.29 and -0.25 with standard deviation of 0.80 and 0.92. The mean infit statistics are 1 for both grades. The mean outfit statistics for grades 4 and 7 are 1.04 and 1.03 respectively.

Figure 7.12.3: MEAP Fall 2012 IRT Difficulty Parameter Distribution for Math

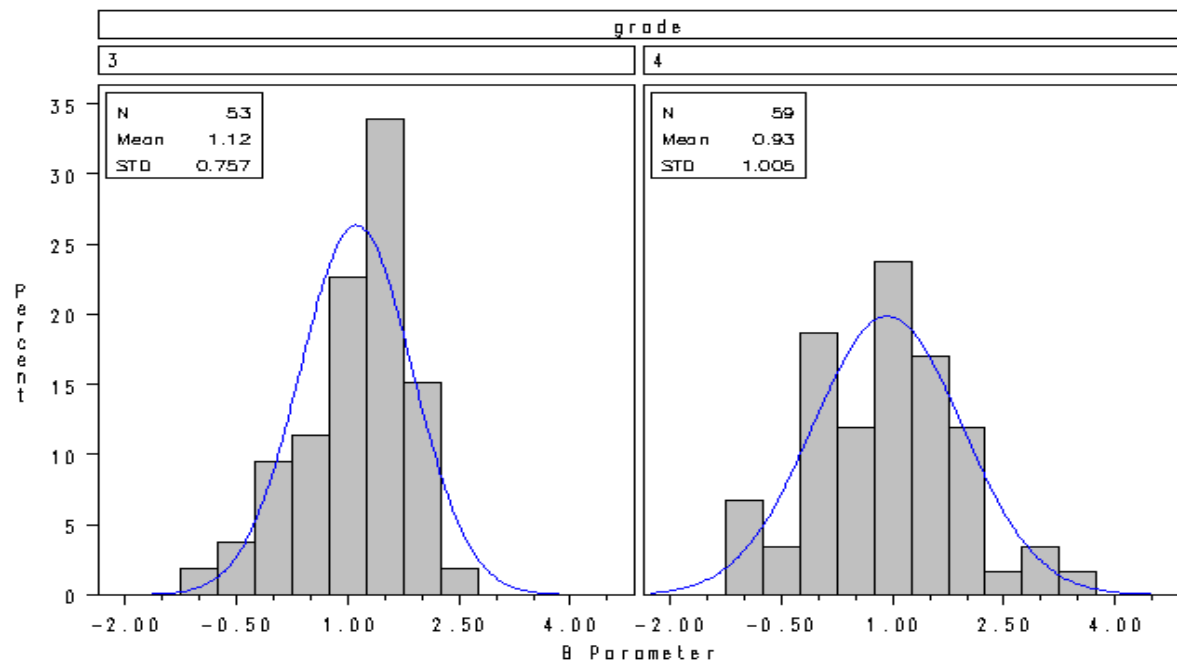


Figure 7.12.3: MEAP Fall 2012 IRT Difficulty Parameter Distribution for Math

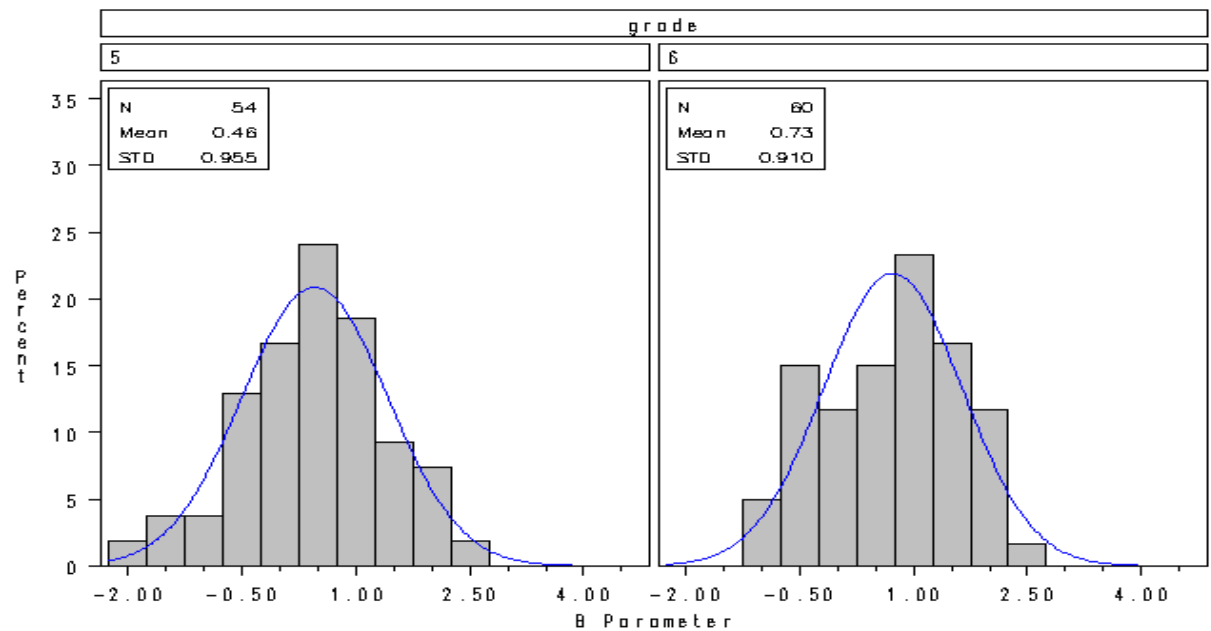


Figure 7.1.2.3: MEAP Fall 2012 IRT Difficulty Parameter Distribution for Math

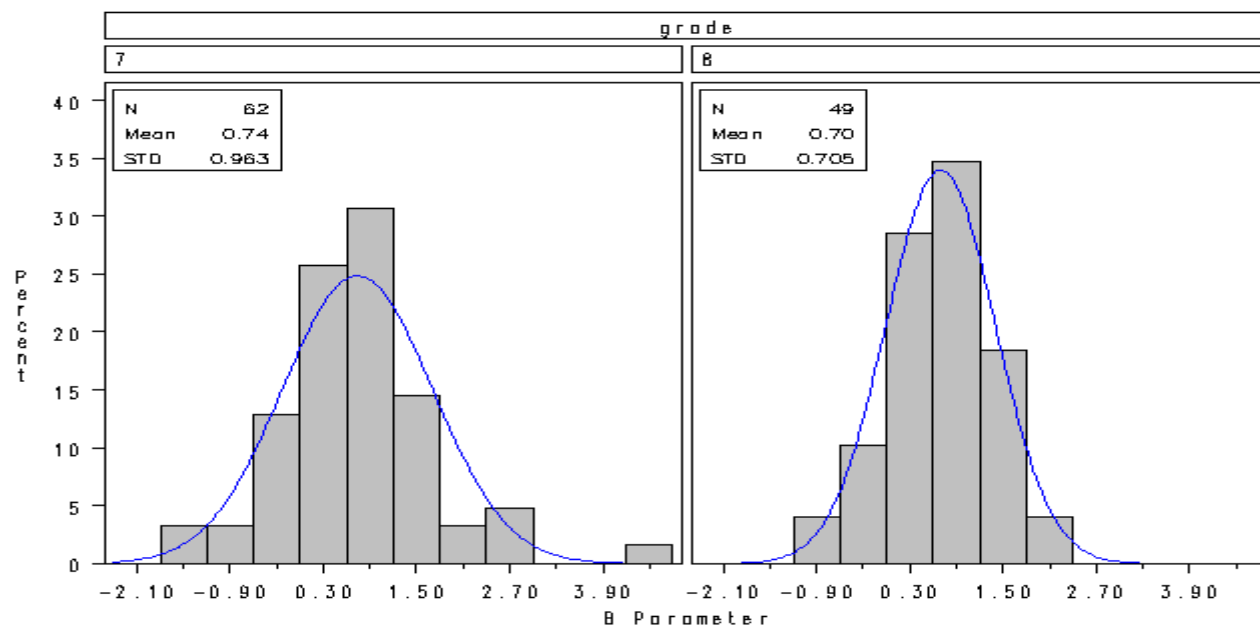


Figure 7.1.2.3: MEAP Fall 2012 IRT Difficulty Parameter Distribution for Math

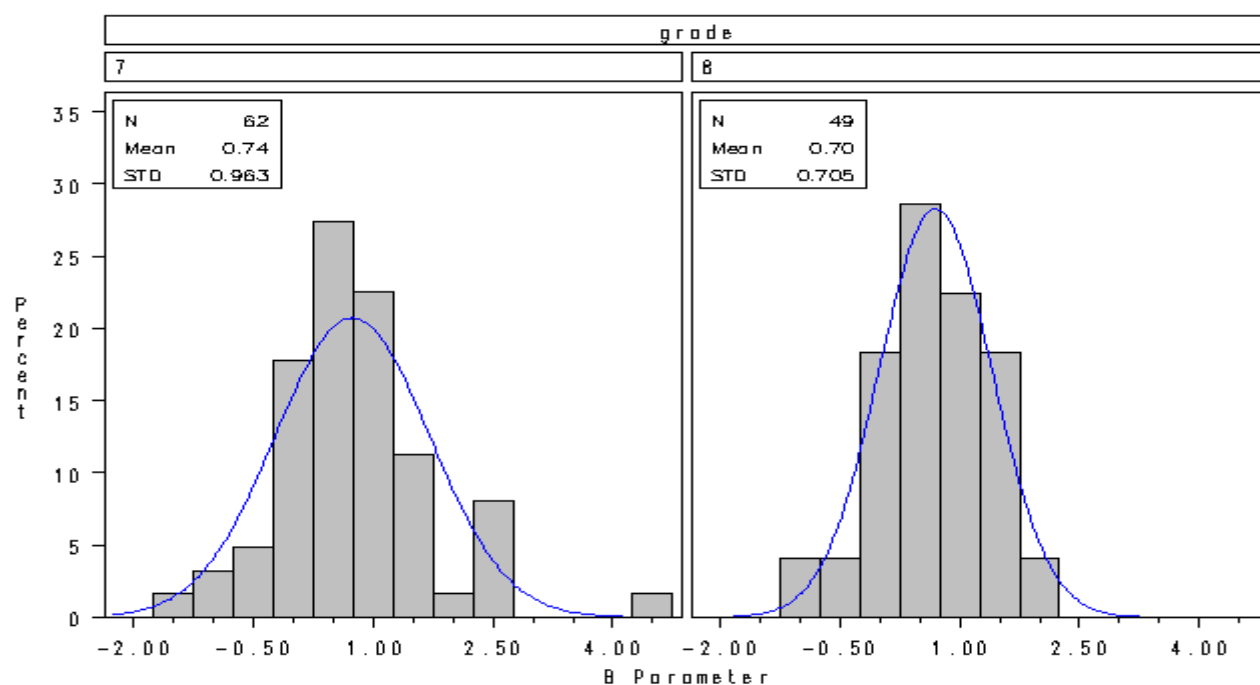


Figure 7.1.2.3: MEAP Fall 2012 IRT Difficulty Parameter Distribution for Reading

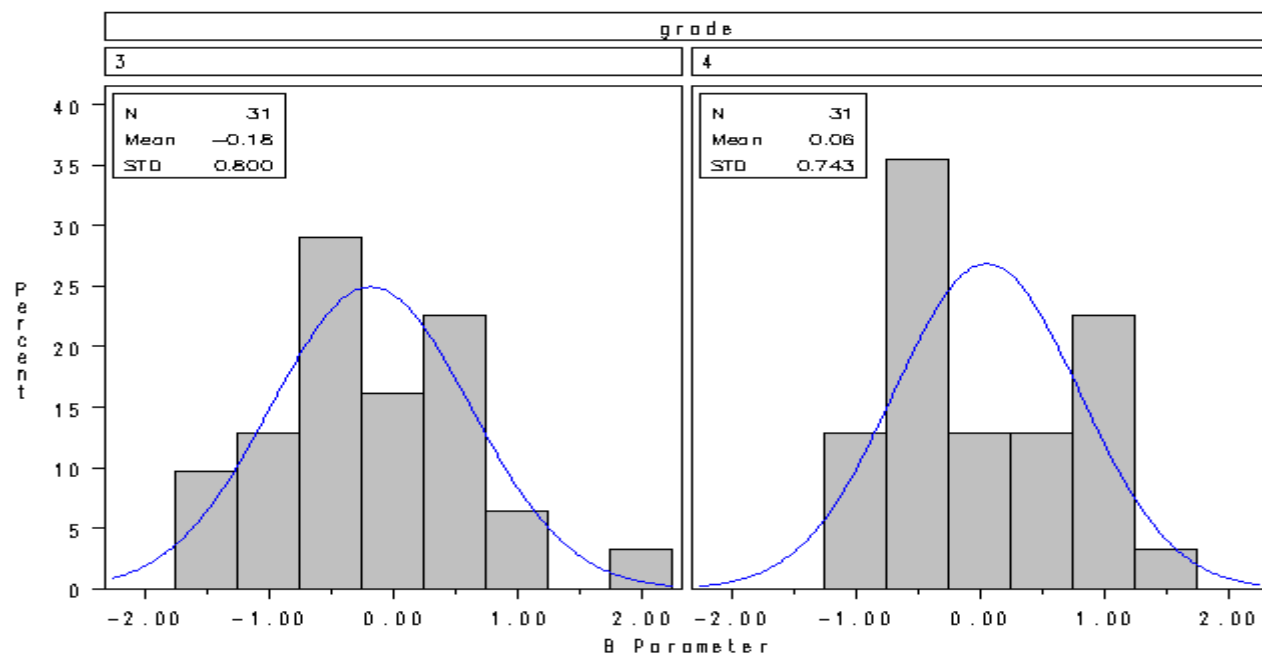


Figure 7.1.2.3: MEAP Fall 2012 IRT Difficulty Parameter Distribution for Reading

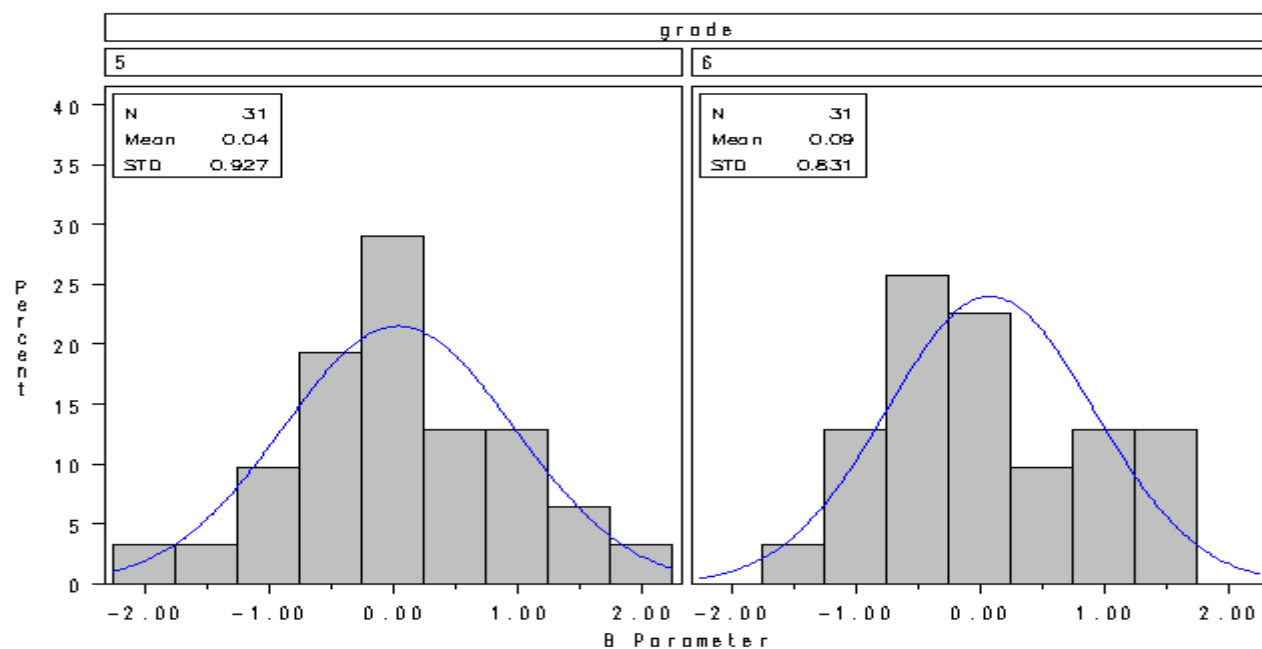


Figure 7.1.2.3: MEAP Fall 2012 IRT Difficulty Parameter Distribution for Reading

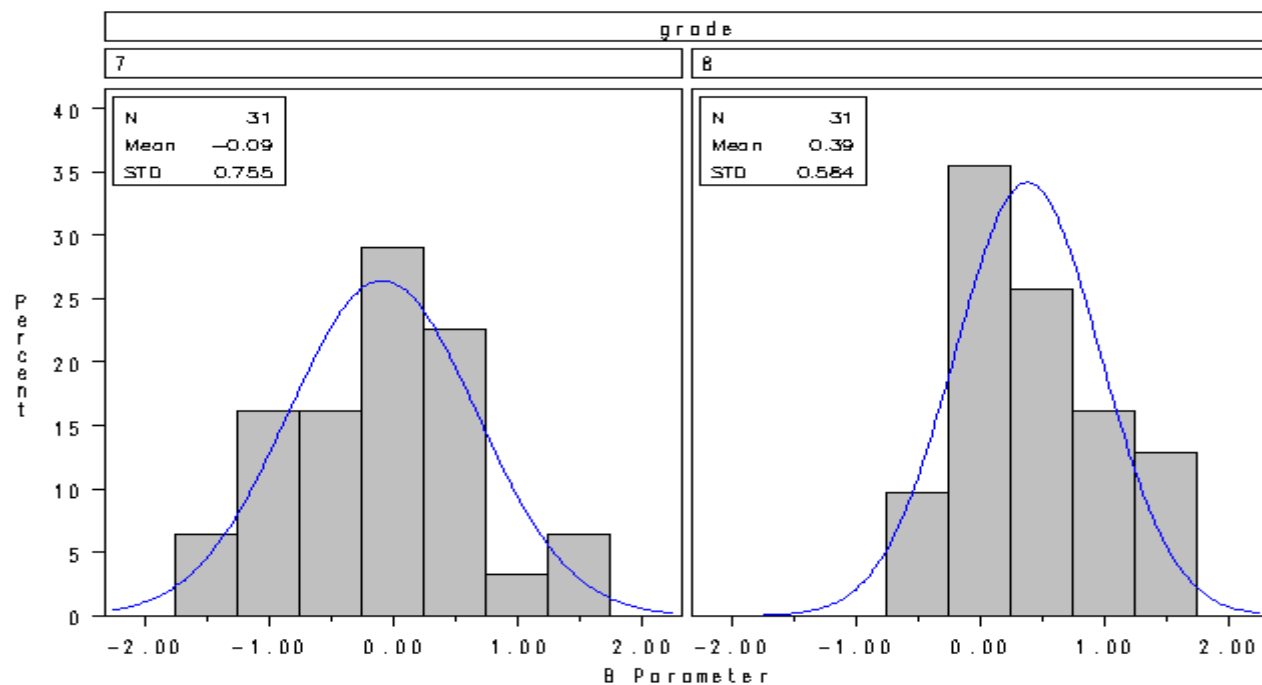


Figure 7.1.2.3: MEAP Fall 2012 IRT Difficulty Parameter Distribution for Science

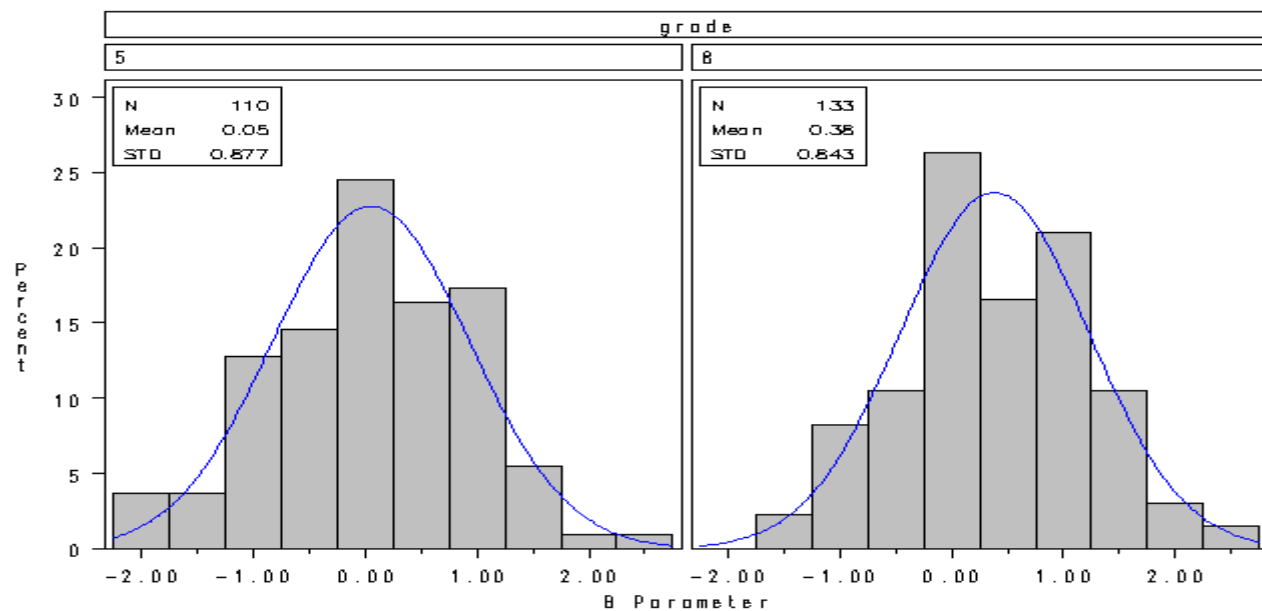


Figure 7.1.2.3: MEAP Fall 2012 IRT Difficulty Parameter Distribution for Social Studies

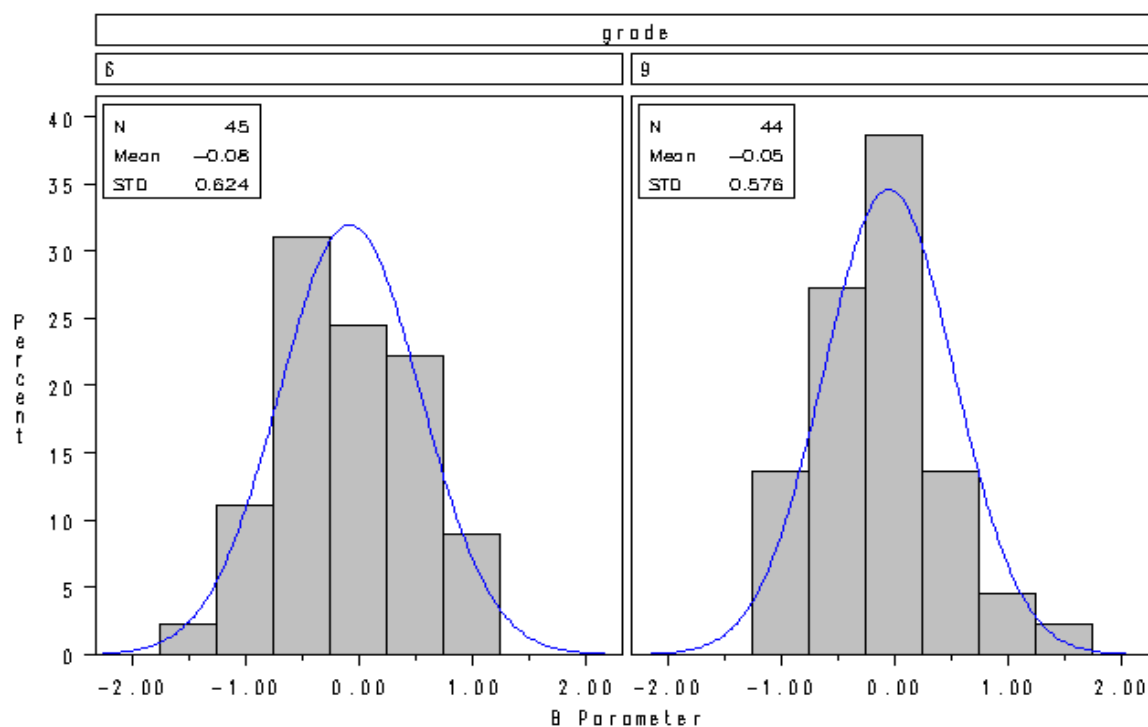
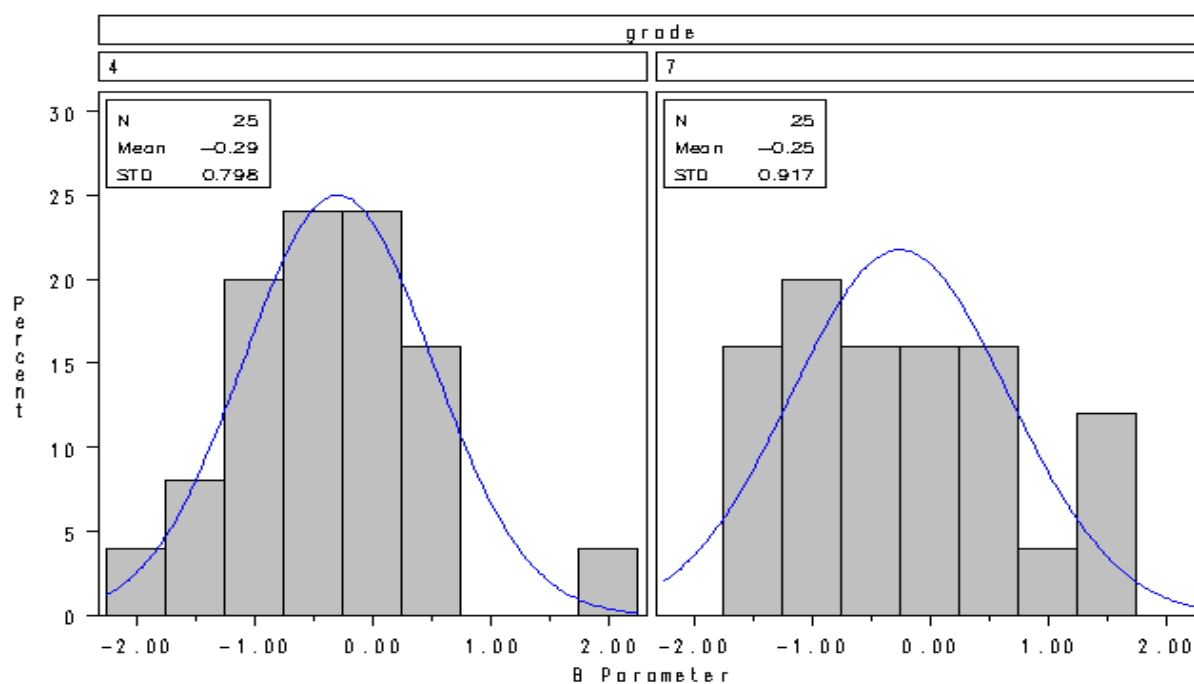
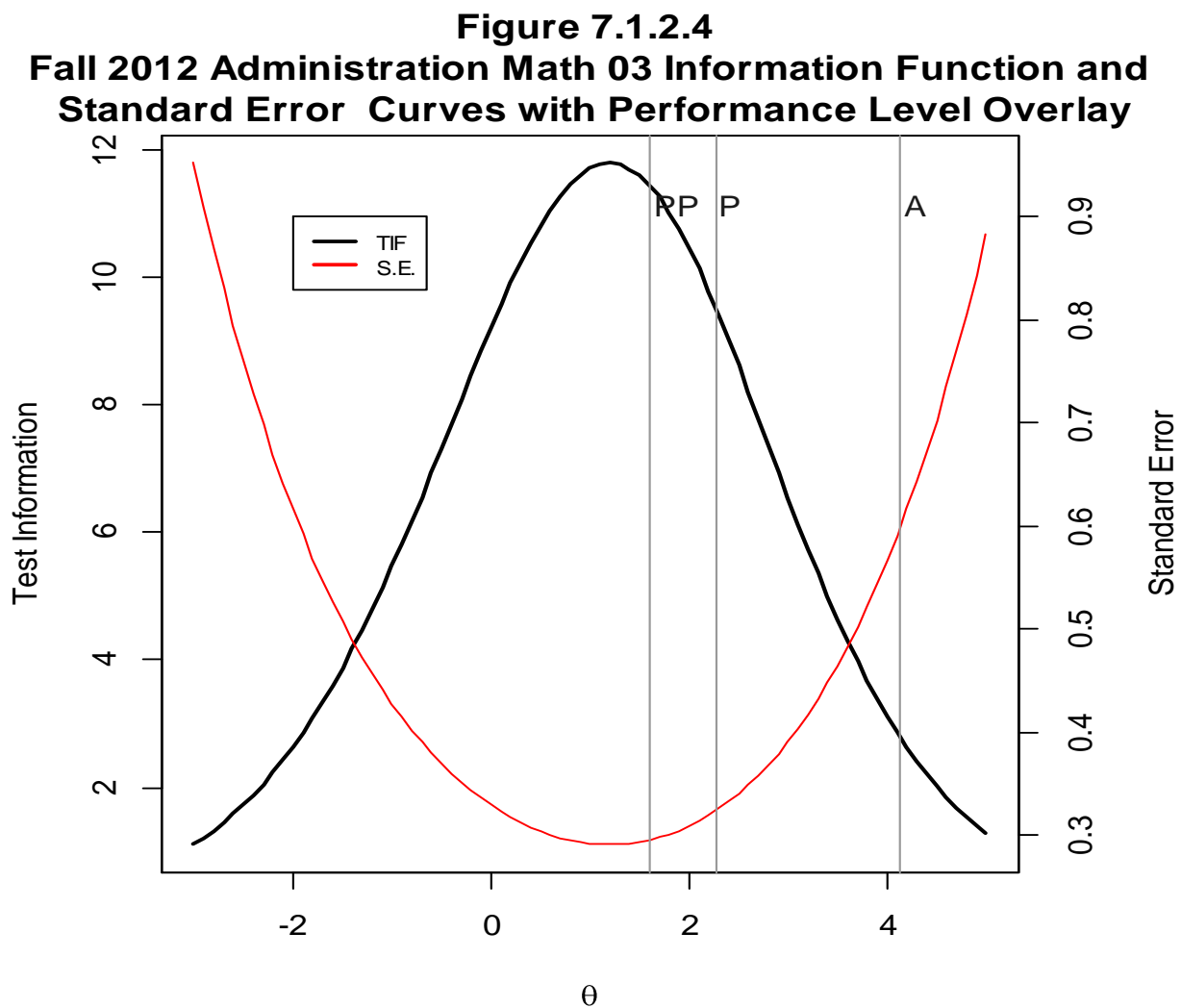


Figure 7.1.2.3: MEAP Fall 2012 IRT Difficulty Parameter Distribution for Writing

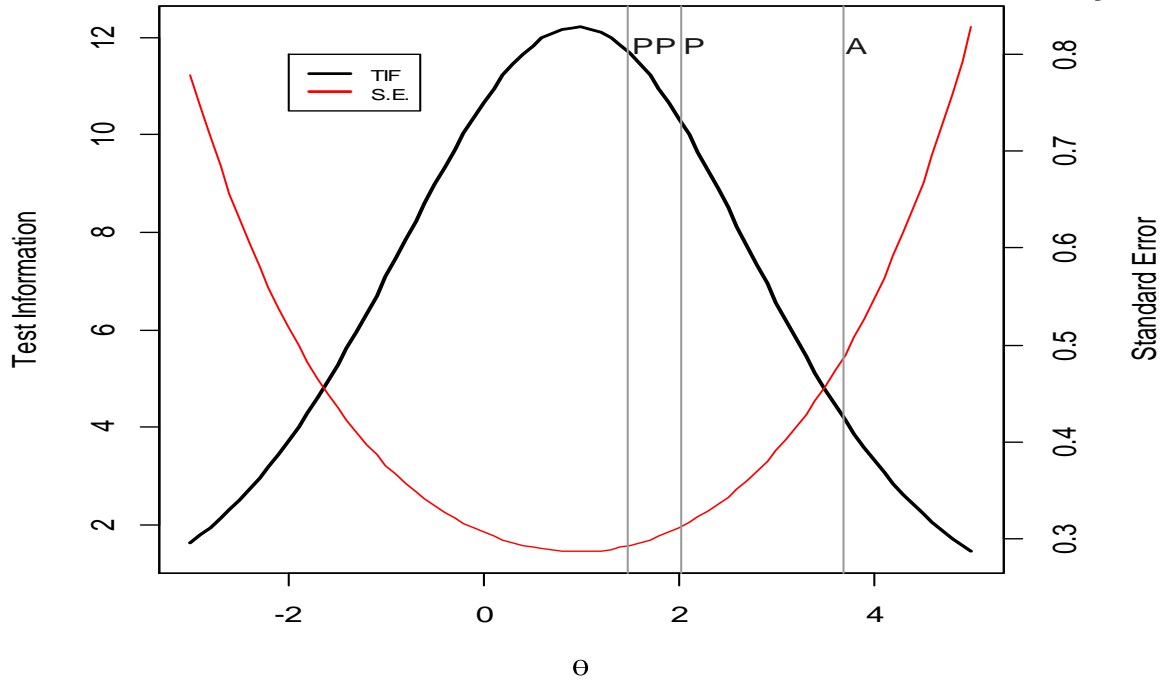


#### 7.1.2.4. Test Information/Standard Error Curves

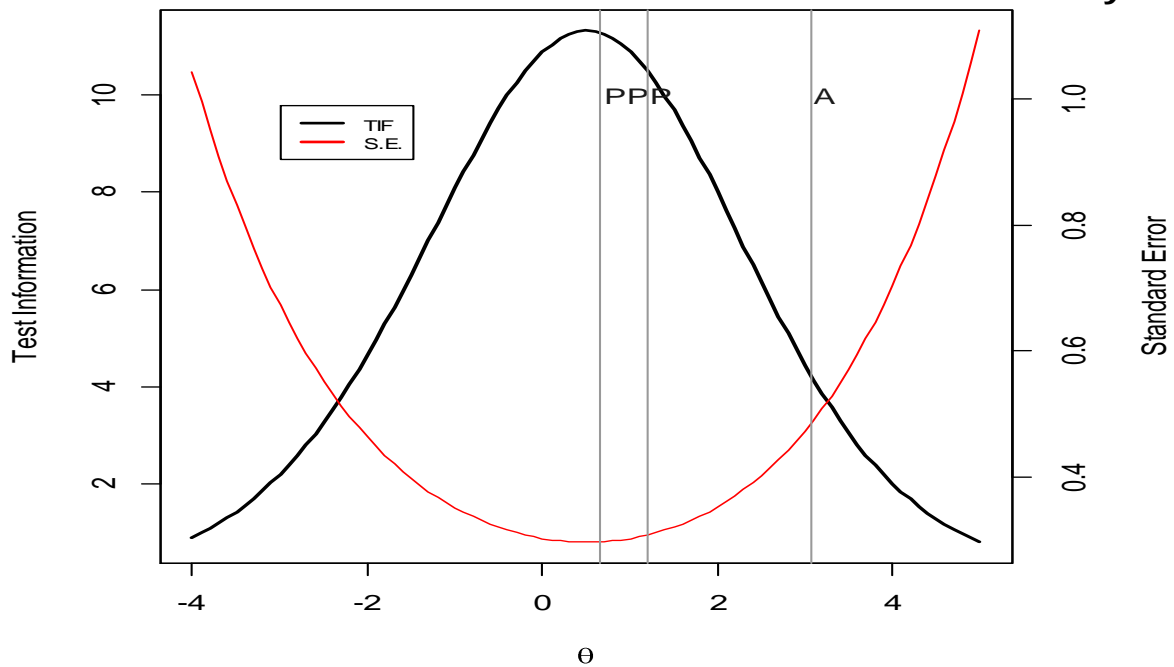
Figures 7.1.2.4 provide the test information function and conditional standard error curves by form, grade, and subject.



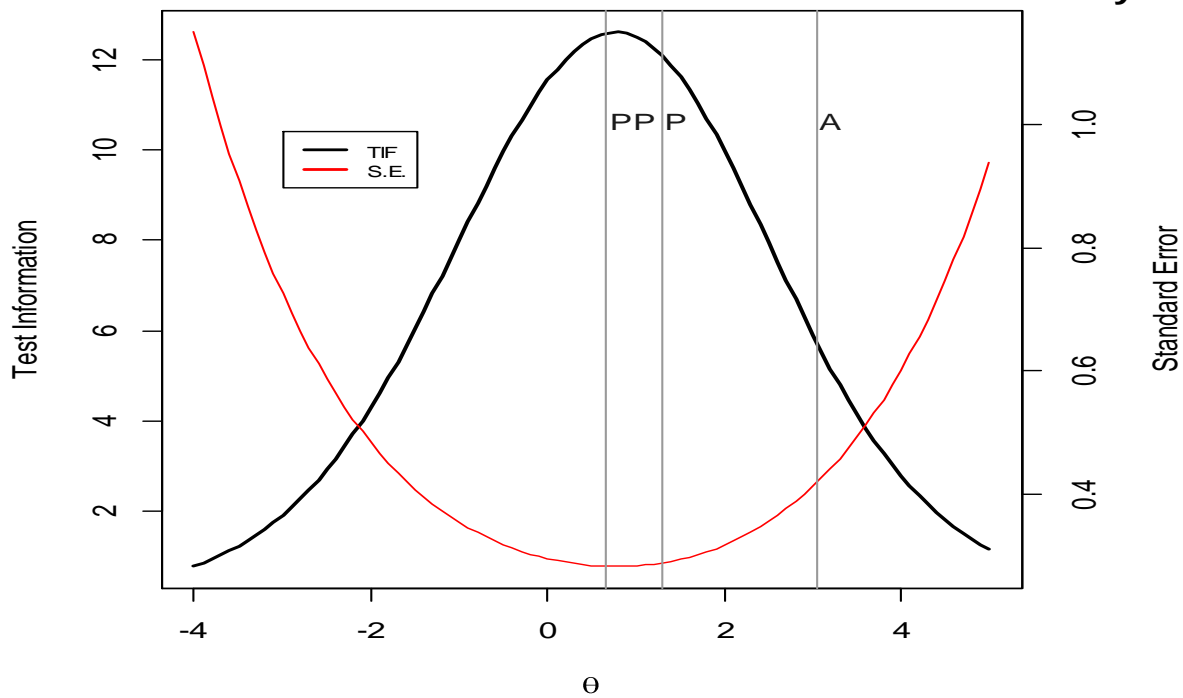
**Figure 7.1.2.4**  
**Fall 2012 Administration Math 04 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



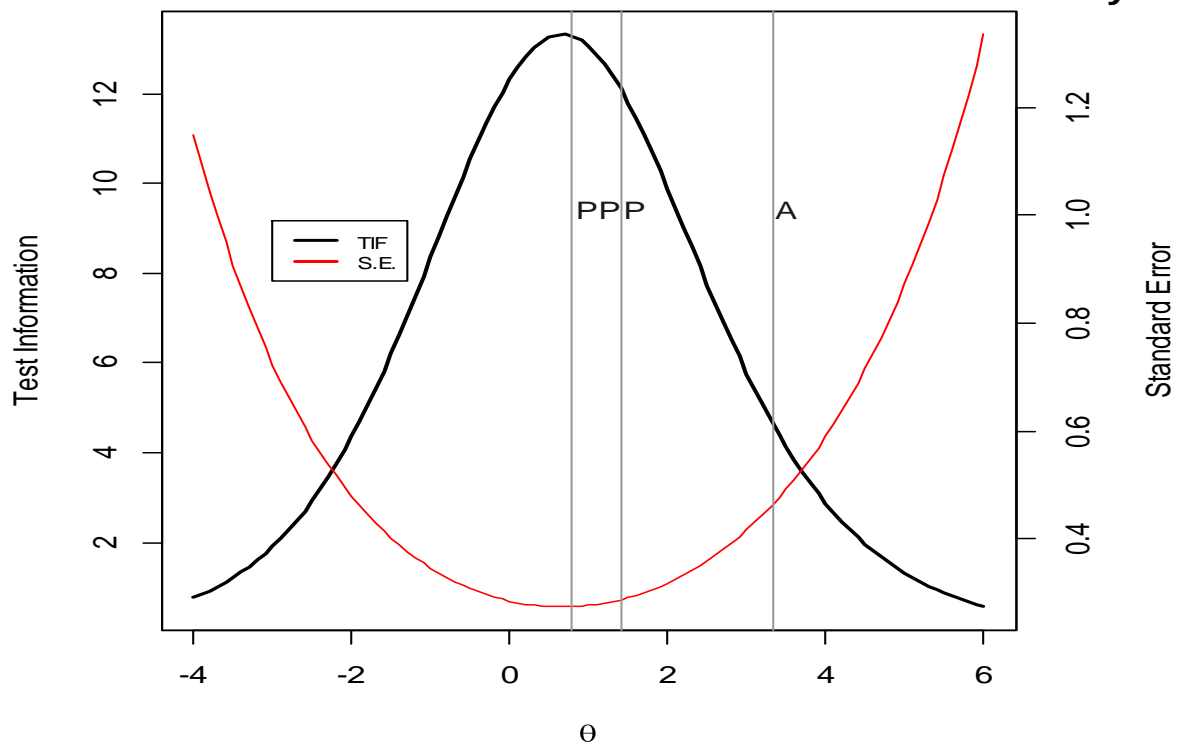
**Figure 7.1.2.4**  
**Fall 2012 Administration Math 05 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



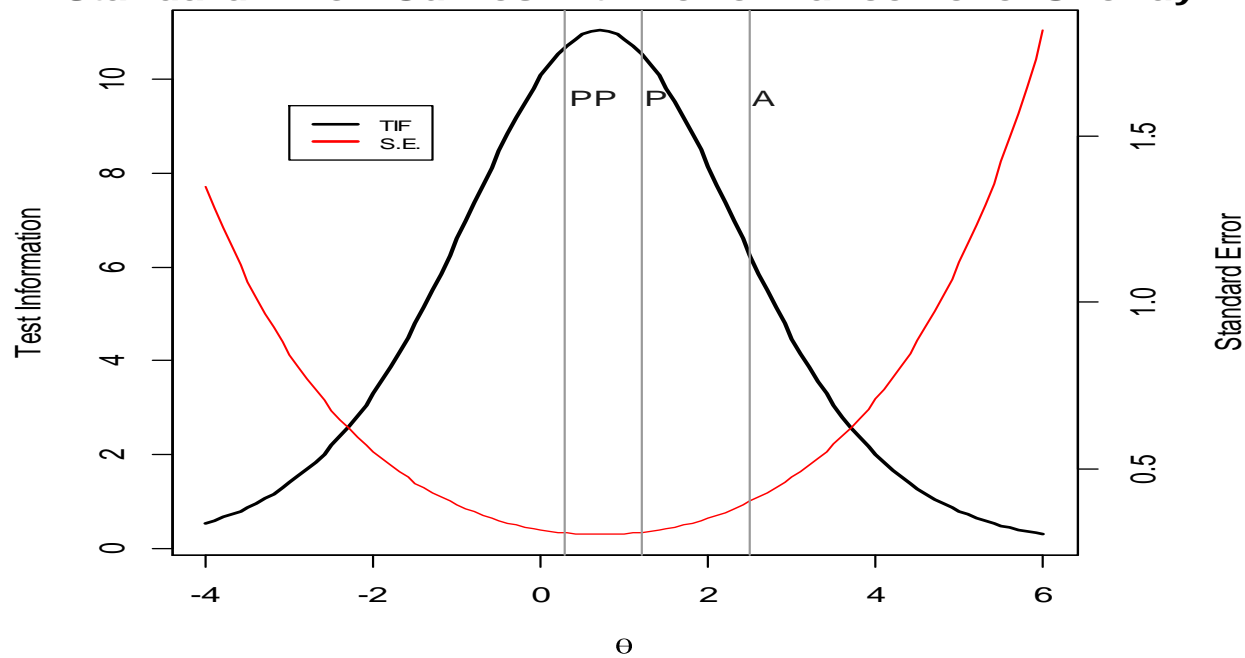
**Figure 7.1.2.4**  
**Fall 2012 Administration Math 06 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



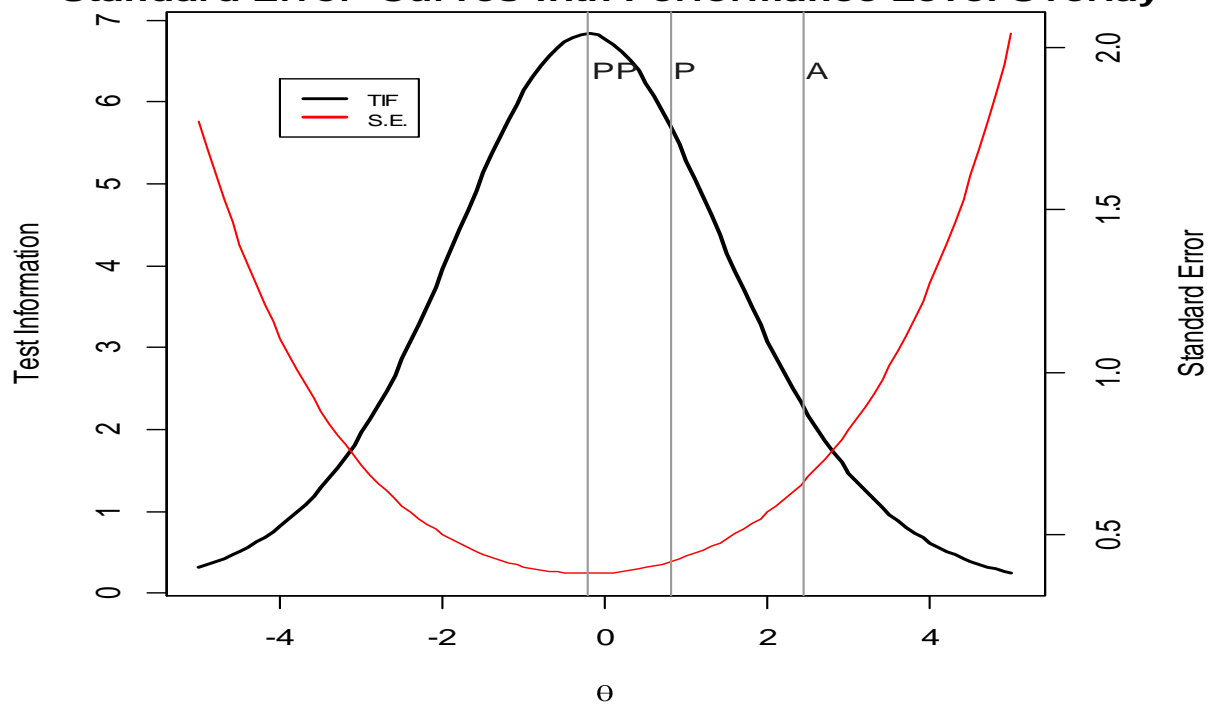
**Figure 7.1.2.4**  
**Fall 2012 Administration Math 07 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



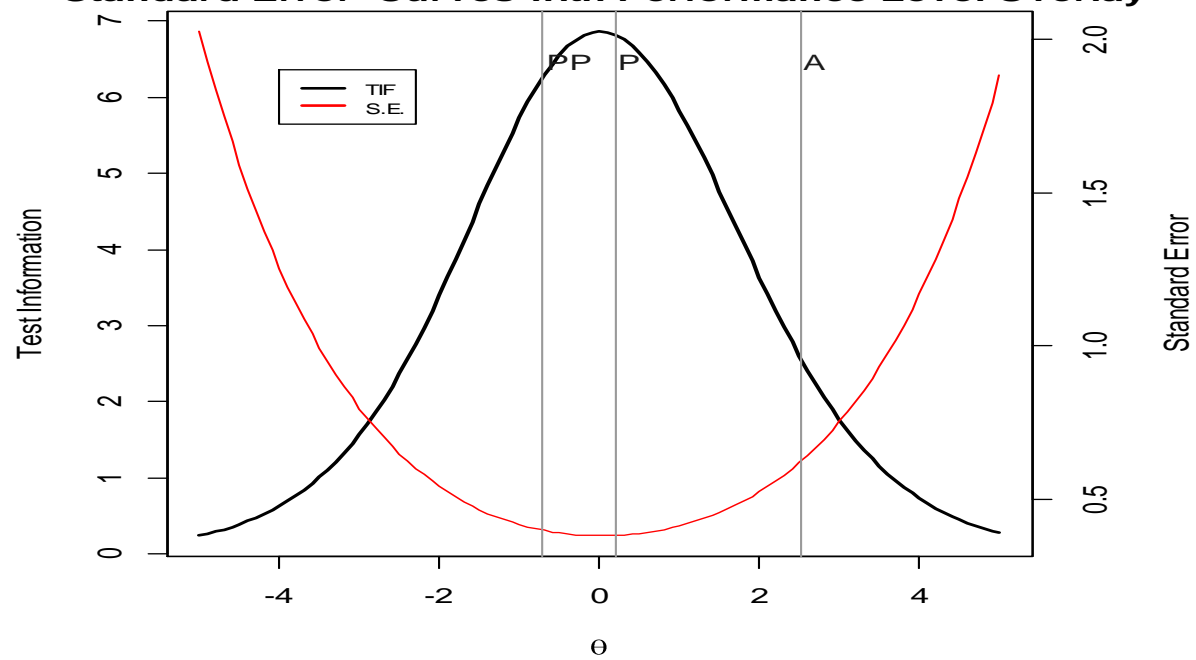
**Figure 7.1.2.4**  
**Fall 2012 Administration Math 08 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



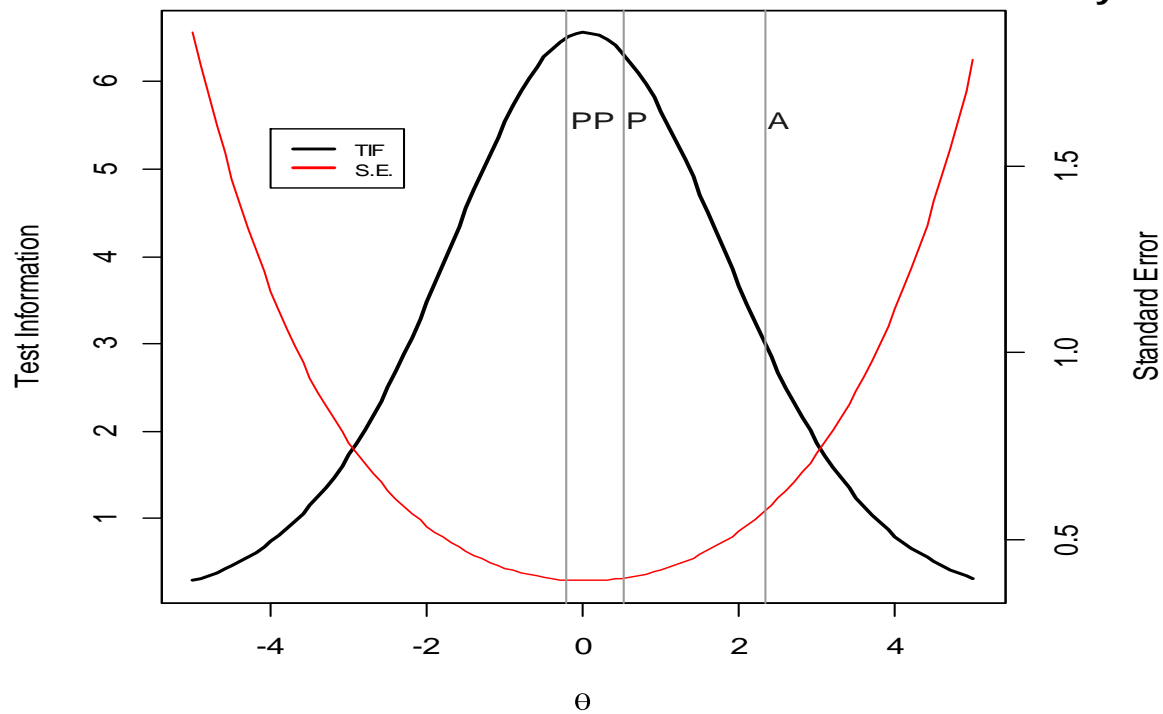
**Figure 7.1.2.4**  
**Fall 2012 Administration Reading 03 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



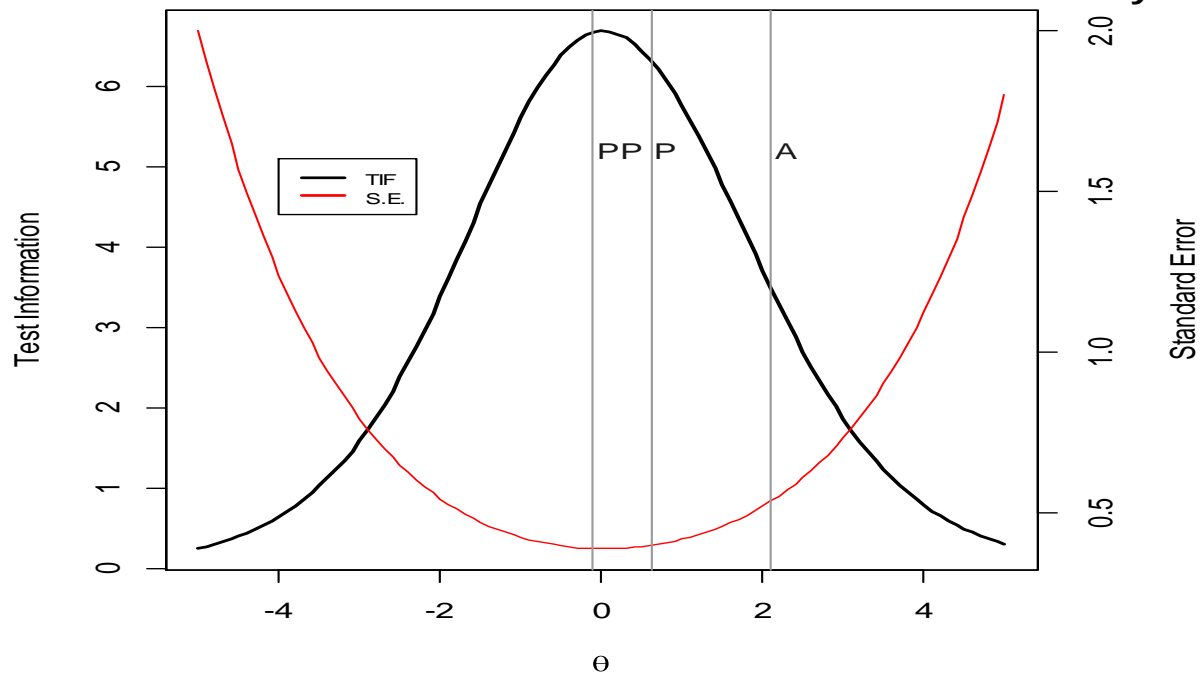
**Figure 7.1.2.4**  
**Fall 2012 Administration Reading 04 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



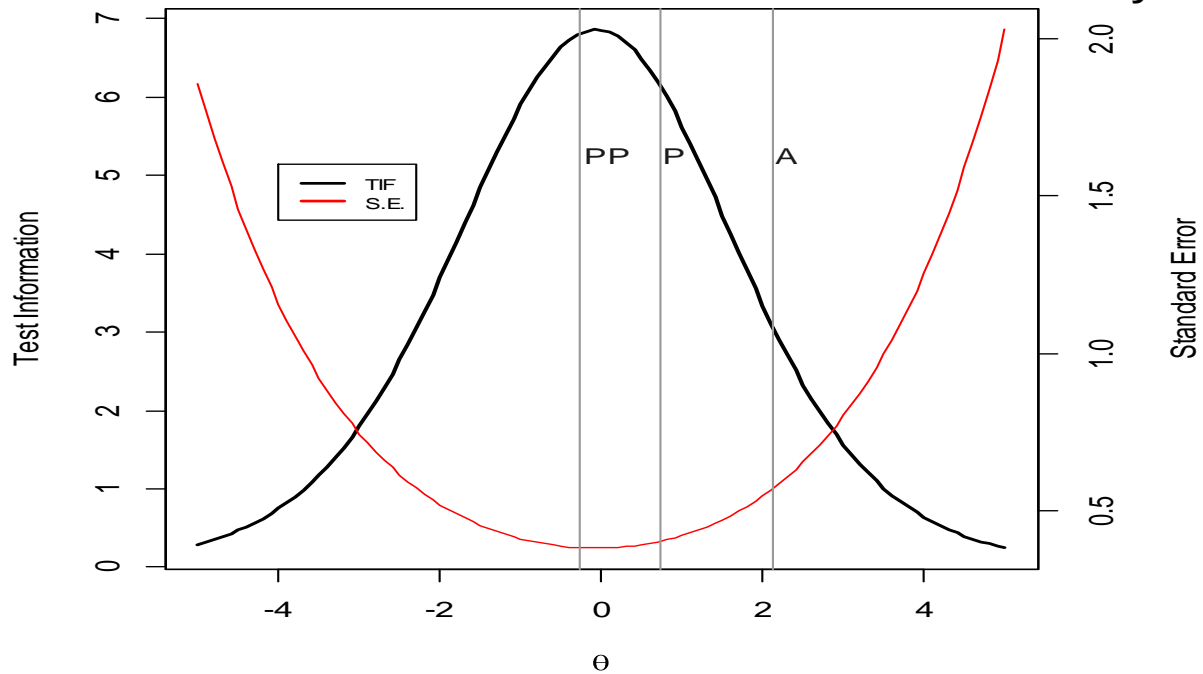
**Figure 7.1.2.4**  
**Fall 2012 Administration Reading 05 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



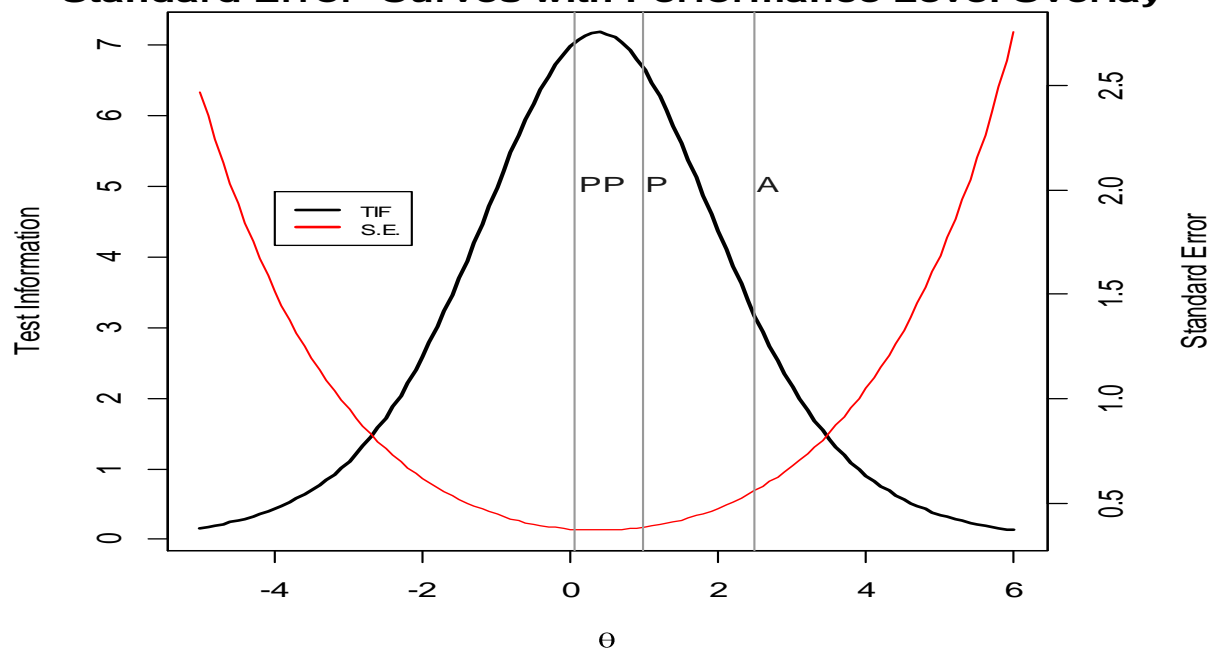
**Figure 7.1.2.4**  
**Fall 2012 Administration Reading 06 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



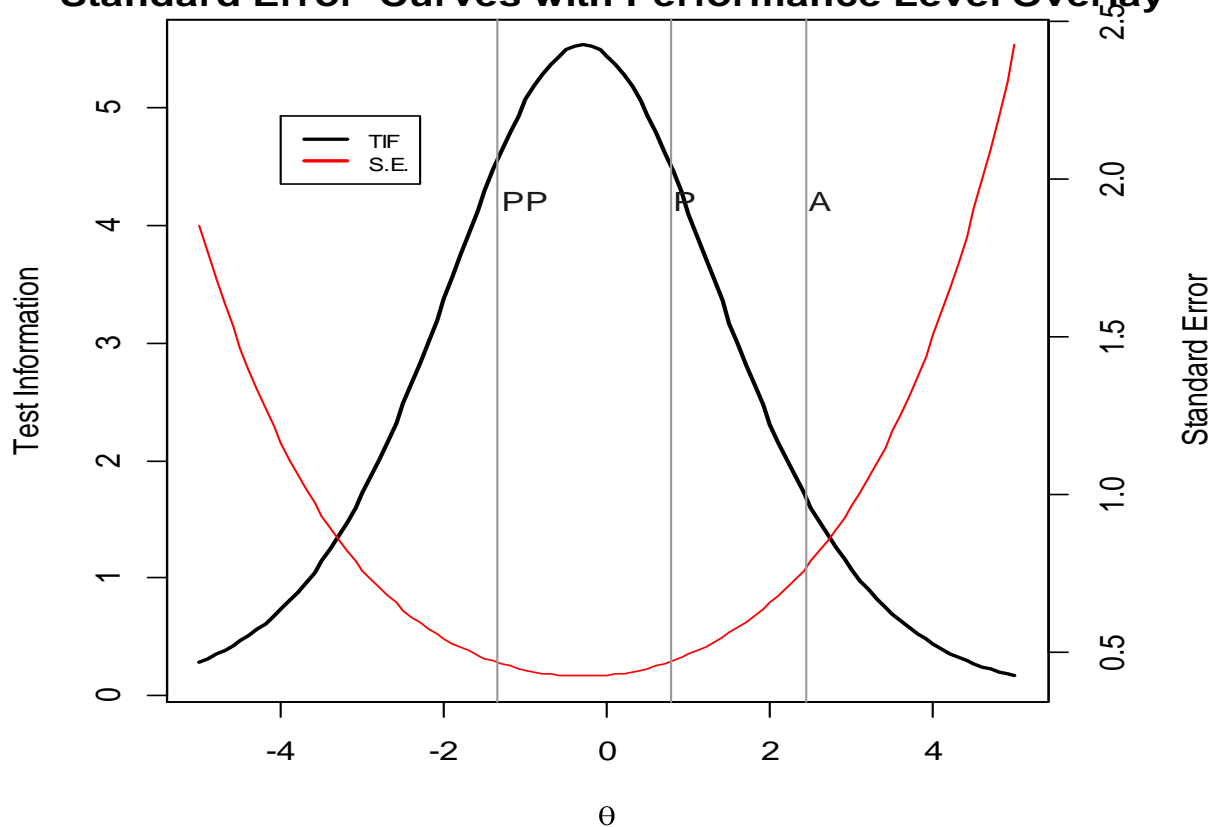
**Figure 7.1.2.4**  
**Fall 2012 Administration Reading 07 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



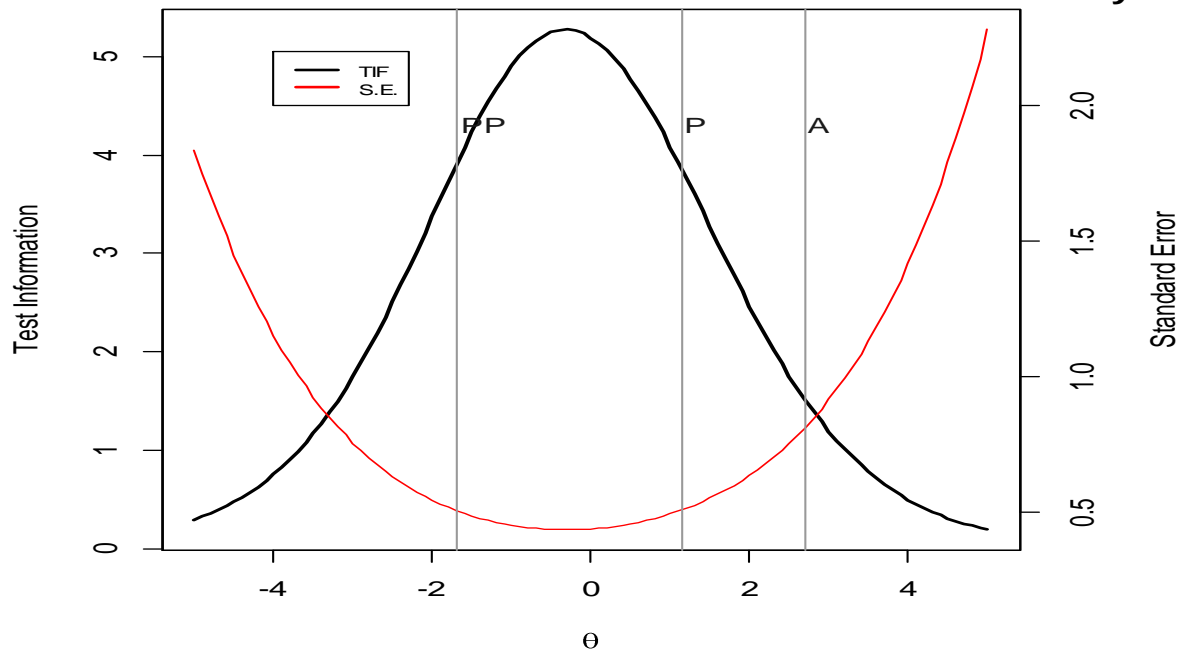
**Figure 7.1.2.4**  
**Fall 2012 Administration Reading 08 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



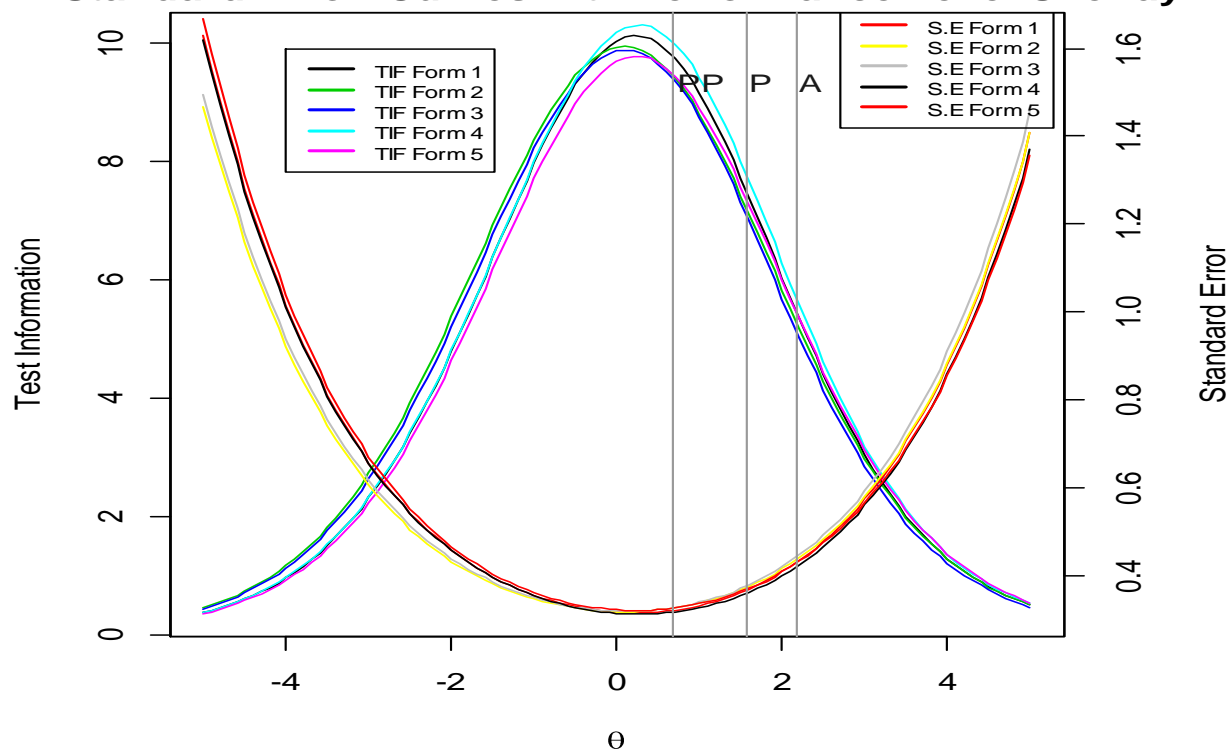
**Figure 7.1.2.4**  
**Fall 2012 Administration Writing 04 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



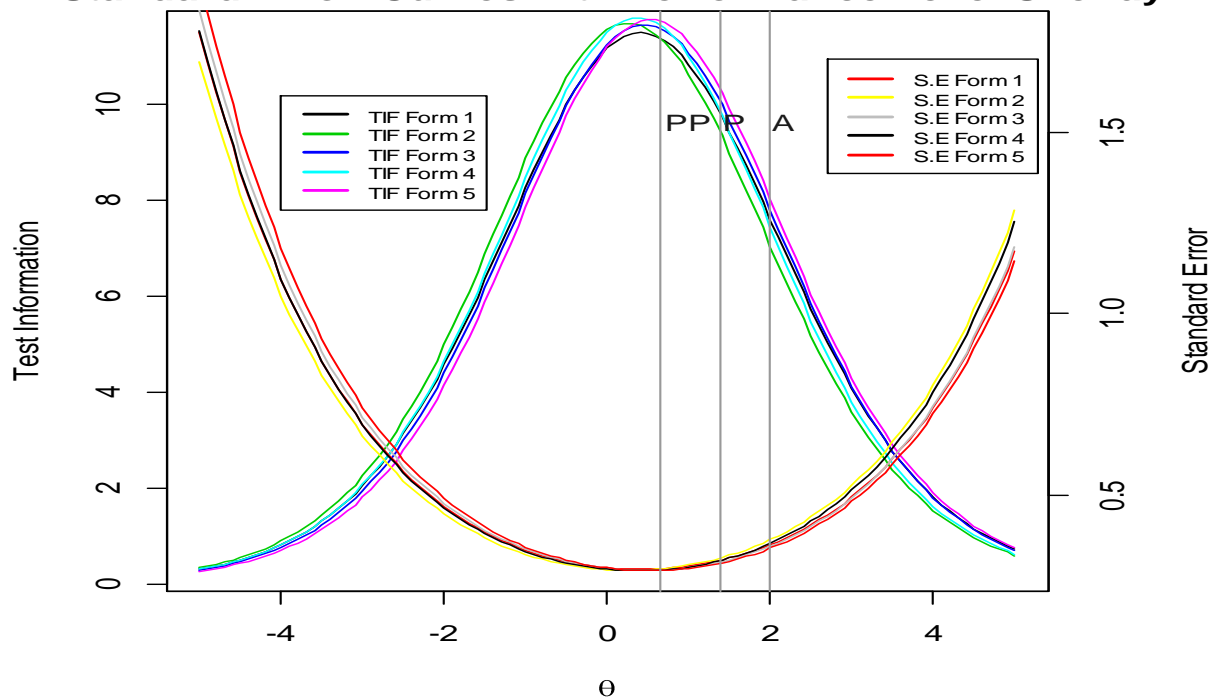
**Figure 7.1.2.4**  
**Fall 2012 Administration Writing 07 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



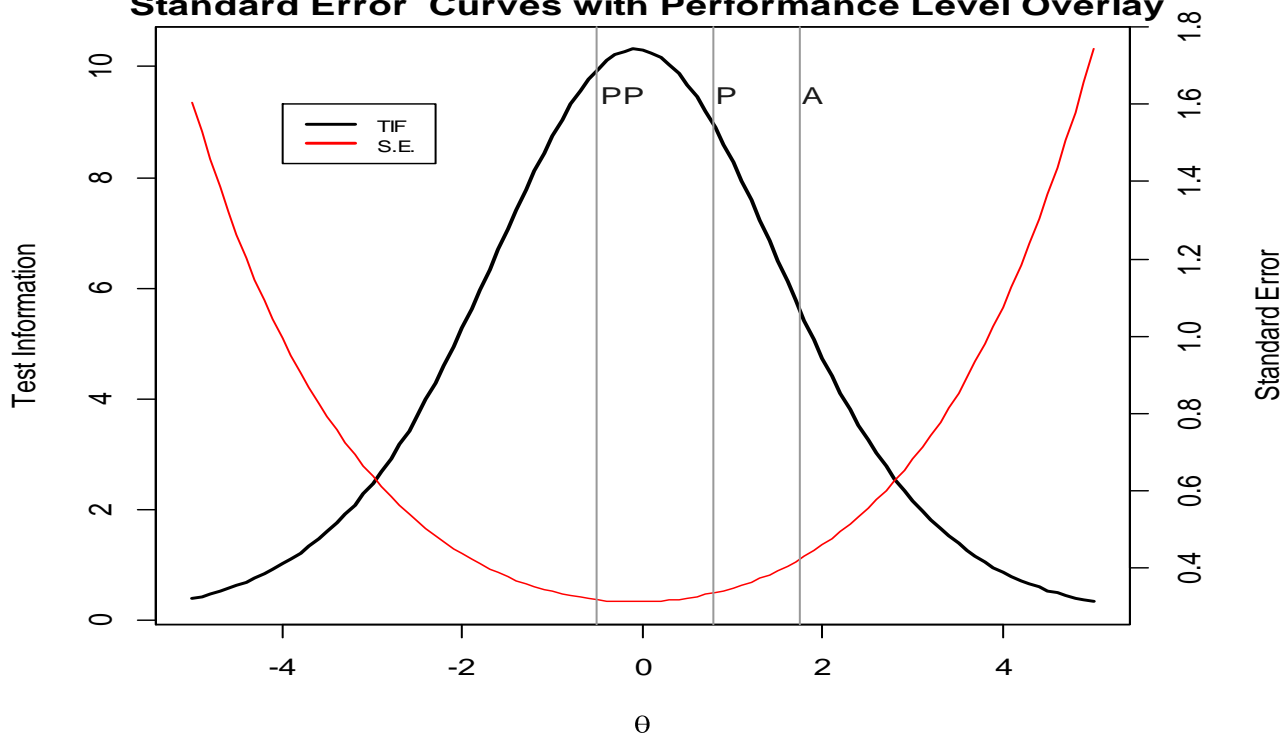
**Figure 7.1.2.4**  
**Fall 2012 Administration Science 05 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



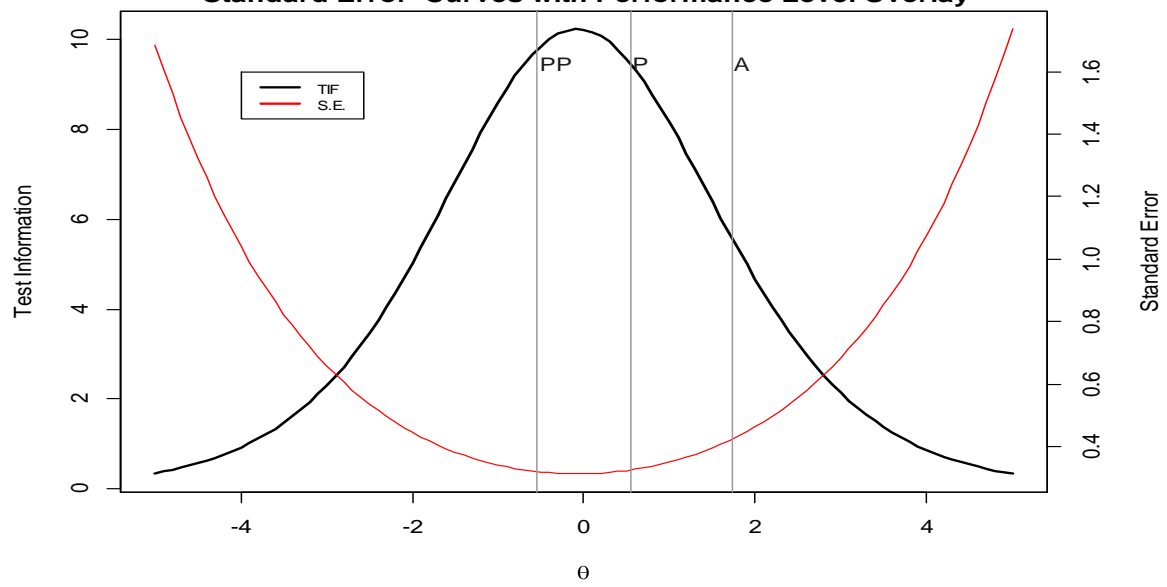
**Figure 7.1.2.4**  
**Fall 2012 Administration Science 08 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



**Figure 7.1.2.4**  
**Fall 2012 Administration Social Studies 06 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



**Figure 7.1.2.4**  
**Fall 2012 Administration Social Studies 09 Information Function and**  
**Standard Error Curves with Performance Level Overlay**



#### 7.1.2.5. Summary of Model Fit Analyses

The tables in *Appendix N* show the summary of the item fit statistics. All subjects have 100 percent infit statistics within (0.5, 1.5) category except for Mathematics Grade 3 (98.11 percent). For outfit statistics 96.23 percent of grade 3, 98.31 percent of grade 4, 98.15 percent of grade 5, 98.33 percent of grade 6, 96.77 percent of grade 7, and 100 percent of grade 8 in mathematics, and 97.92 of grade 5 all forms, 98.11 of grade 8 form1 and 5 in science, and 100 percent of reading, social studies, and writing was between 0.5 and 1.5. Overall, based on the infit and outfit statistics, all items fit well for the Rasch model.

### 7.2. Scale Scores

#### 7.2.1. Description of the MEAP Scale

There are four performance levels for MEAP: Not Proficient, Partially Proficient, Proficient, and Advanced. Three cut-scores are needed to define the four performance levels. For MEAP Grades 3-8, the BAA has decided to set the cut score for “Proficient on Michigan Standards” for a given grade to be X00 for grade X, such that 300 is the equivalent Proficient scale score for grade 3, 400 for grade 4, 500 for grade 5, 600 for grade 6, 700 for grade 7, and 800 for grade 8. To set a scale for each given grade, either two scale score points need to be set or one scale score point and the variability need to be set. To set the MEAP scale, a common standard deviation across grades is set at 25.

#### 7.2.2. Identification of the Scale, Transformation of IRT Results to MEAP Scale

Tables 7.2.2.1 show the predetermined target  $\theta$ s (Met Michigan standard) from the standard setting activities for MEAP tests and the scale scores. The MEAP scale scores were created from the following formula:

$$SS = SS_{met} + \frac{\sigma_{SS}}{\sigma_{\theta}}(\theta - \theta_{met})$$

Where  $SS$  indicates the scale score, and  $\theta$  indicates the  $\theta$  value. The values of each variable (except for  $\theta$  and  $SS$ ) are given in the following table for every grade and subject:

**Table 7.2.2.1**  
**Thetas from Standard Setting and Scale Scores for MEAP Grades 3-8**

Subject	Grade	$\sigma_{\theta}$	$\sigma_{SS}$	$\theta_{met}$	$SS_{met}$
Mathematics	3	1.322	25	0.430	300
	4	1.123	25	0.436	400
	5	0.945	25	-0.013	500
	6	0.990	25	0.144	600
	7	1.005	25	0.201	700
	8	0.961	25	-0.005	800
Reading	3	1.107	25	-0.302	300
	4	1.010	25	-0.540	400
	5	1.041	25	-0.332	500
	6	1.078	25	-0.167	600
	7	0.993	25	-0.117	700
	8	1.069	25	0.271	800
Science	5	0.875	25	-0.232	500
	8	0.918	25	-0.301	800
Social	6	1.053	25	-0.262	600
Studies	9	0.963	25	-0.493	900
Writing	4	1.402	25	0.781	400
	7	1.284	25	1.083	700

### 7.2.3. Scale Score Interpretations and Limitations

Because the scale scores associated with the On-Grade MEAP are not a vertical scale, care must be taken before any interpretation of individual scale score differences between grades is made. It is important to note, however, that only the passing achievement level is constant across grades or subjects. Comparisons of scale scores across subjects are even more suspect. In general, achievement levels are the best indicators for comparison across grade or subject. For future years, when a vertical scale is developed, more meaningful comparisons across grades will be possible.

The scale scores can be used to direct students needing remediation (i.e., students falling below Basic level), but scale score gain comparisons between individual students are not appropriate. It is acceptable to compare gain scores for groups of students, because measurement precision is increased when scores are aggregated.

Because scale scores and number correct scores are on two distinct score metrics, users should be cautioned against over-interpreting differences in scale scores. As a hypothetical example for grade 4 mathematics, a student near the middle of the scale score distribution might change his or her scale score value of only four points (for example, from 400 to 404) by correctly answering two additional multiple choice questions. However, a student near the top of the scale score distribution may increase his or her scale score of 35 points with two additional questions correctly answered (for example, from 500 to 535). A similar phenomenon may be observed near the bottom of the score scale.

The primary function of the scale score is to be able to determine how far students are from the various proficiency levels without depending upon the changing raw scores. Additionally, schools will use the scale scores in summary fashion for comparisons of program evaluations across the years. For example, it is valid to compare the average grade 5 scale score in mathematics from one year to the average of the previous year. Interpretations of why the differences exist will depend on factors specific to individual schools.

#### 7.2.4. Upper and Lower End Scaling

MEAP scale scores are a linear transformation from  $\theta$  to scale scores, and there is no adjustment for the upper and lower scale scores. Table 7.2.4.1 shows the scale score ranges for all MEAP assessments.

**Table 7.2.4.1**  
**Scale Score Ranges for MEAP Fall 2012**

<b>Subject</b>	<b>Grade</b>	<b>Lowest SS</b>	<b>Highest SS</b>	<b>Range</b>
MA	03	208	416	208
MA	04	283	539	259
MA	05	363	668	305
MA	06	470	769	299
MA	07	572	863	293
MA	08	668	950	289
RD	03	188	423	235
RD	04	283	537	254
RD	05	385	630	245
RD	06	490	730	240
RD	07	574	826	252
RD	08	688	921	233
SC	05	350	668	318
SC	08	668	971	303
SS	06	481	729	248
SS	09	778	1046	268
WR	04	247	513	266
WR	07	531	809	278

## **CHAPTER 8: EQUATING**

### **8.1. Rationale**

To maintain the same performance standards across different administrations, all tests must have comparable difficulty. This comparable difficulty is maintained from administration to administration at the total test level and, as much as possible, at the reporting strand level. A pre-equating procedure is applied on the MEAP. This equating design ensures that the level for any performance standard established by the MEAP on the original test is maintained on all subsequent test forms.

Each test form consists of base items and field-test items. Base items are those items that are the same across all test forms within each subject and grade and count toward a student's score. Field-test items are those being administered for the first time to gather statistical information about the items. They are also used for linking to future forms, for some administrations, and for generating school level scores. These items do not count toward an individual student's score.

Technically, the 2012 administration of the grade 3-9 MEAP assessments were equated to Fall 2011. The details in this chapter explain the procedure that is used for equating 2012 and future forms to the 2011 scale.

### **8.2. Pre-equating**

In the pre-equating process, a newly developed test is linked to a set of items that were previously used on one or more test forms. In this way, the difficulty level of the newly developed test can be equated through the linking items to previously administered tests. This procedure is known as common item equating. For the Fall 2012 administration, each new assessment is constructed from a pool of items that have been equated back to the 2011 test form.

#### **8.2.1. Test Construction and Review**

Test construction and review for the Fall 2012 MEAP were discussed in detail in section 2.8. Design of test forms and the item selection process were presented in section 2.8.

#### **8.2.2. Field-Test Items**

Once a newly constructed item has survived committee review and is ready for field-testing, it is embedded among the base-test items in a test booklet. The base-test items count toward the individual student's score. For example, on the MEAP grade 5 reading test for a particular administration, there are 5 forms containing the same base-test items. However, each form would also contain 24 unique field-test items, which vary by form (The field-test items do not count toward an individual student's score and may be used as equating or linking items to past or future tests).

For 2012 MEAP administration, a stratified form assignment design was used, to balance form distribution across various demographics. This form assignment procedure provides a diverse sample of student performance on each field-test item. In addition, because students do not know which items are field-test items and which items are base-test items, no differential motivation effects are expected. To control for fatigue and start-up effects, all field-test items are placed in similar positions on each test form.

### **8.2.3. Within-Grade Equating**

#### **8.2.3.1. Description of Procedures Used to Horizontally Equate Scores from Various Forms at the Same Grade Level**

Once the statewide data file has been edited, calibrations are performed on all base-test items regardless of the test form. The WINSTEPS (Linacre and Wright, 2004) program is used to estimate the Rasch item difficulty (RID) parameters and the constructed response task (step) parameters for the MEAP tests. Using pre-equated values from 2011 administration, item parameters are scaled using post equating procedure described in section 7.1.2.1.

#### **8.2.3.2. Item and Test Statistics on the Equated Metric, Test Information/Standard Error Curves, and Ability Distributions**

The tables in *Appendices M and N*, and Figure 7.1.2.4 provided the item and test statistics on the equated metric, test information/standard error curves, and ability distribution. More detailed information can be found in sections 7.1.2.3 and 7.1.2.4.

### **8.3. Vertical Equating**

Vertical linking was not created during year 2012. To make test development more efficient, vertical linking design was dropped from test design.

### **8.4. Ability Estimation**

The item and step difficulties from the anchored calibration run are used in conjunction with actual student performance to obtain Rasch ability estimates for each possible raw score value. The generation of this raw score-to-Rasch ability is accomplished through a variation of the fundamental formulas in Rasch model measurement (Wright, 1977, p. 110) using the WINSTEPS computer program.

The process for obtaining Rasch ability estimates for the MEAP tests is complicated. This is because of the combination of dichotomously scored multiple choice items and polytomously scored constructed-response tasks. The procedure outlined by Masters and Evans (1986) adapted for this purpose is summarized in the following paragraphs.

First, using the item and step difficulty estimates from the anchored calibration run, follow this procedure:

For each raw score  $R$ , begin by assuming  $A_R = B_R$ , where  $A_R$  is the updated estimate from the procedure and  $B_R$  is the initial ability estimate and is equal to:

$$\ln \times [R / (T - R)] ,$$

where T is the maximum score possible:

$$T = \sum_{i=1}^L m_i ,$$

where L is the number of items on the test.

For each item  $i$ , calculate the following:

$$Q_{XR} = \exp \left( XA_R - \sum_{j=1}^X \delta_{ij} \right) ,$$

where  $X = 1, 2, \dots, m_i$ ;  $\delta_{ij}$  are item parameters, and

$$P_{XR} = \frac{Q_{XR}}{\left( 1 + \sum_{K=1}^{m_i} Q_{KR} \right)} ,$$

where  $X = 1, 2, \dots, m_i$ , and

$$Y_i = \sum_{K=1}^{m_i} KP_{KR} ,$$

and

$$Z_i = \sum_{K=1}^{m_i} K^2 P_{KR} .$$

The improved ability estimate is then calculated by

$$B_R = \frac{A_R + \left( R - \sum_{i=1}^L Y_i \right)}{\sum_{i=1}^L (Z_i - Y_i^2)} .$$

This procedure represents one cycle or iteration in the calculation of the ability estimate.  $B_R$  from this procedure is then used as the next cycle's initial ability estimate (replacing  $A_R$ ) and the cycle continues. If the absolute difference between the initial and final ability estimate is less than 0.01, the cycles are terminated and the current  $B_R$  value is used as the ability estimate (Masters and Evans, 1986, p. 365).

## **8.5. Development Procedures for Future Forms**

### **8.5.1. Equating Field-Test Items**

Field-test items are equated using WINSTEPS and anchored on all operational test items by grade and content area. This anchored calibration produces results that all of the field-test items are on the same scale as the base test.

The base test for future administrations will be equated to the 2012 scale by anchoring the common item difficulty values to those obtained in 2012 and allowing WINSTEPS to estimate the new item and task/step difficulty values to this anchored scale. The result is a base-test form with item and task/step difficulty values on the same scale as the original form administered in 2005.

### **8.5.2. Item Pool Maintenance**

Bureau of Assessment and Accountability (BAA) implemented the statistical programming required to populate the MEAP item bank. Detailed description of the item bank is in Chapter 2, section 2.7.3.

BAA psychometrician completed item level statistics for the Fall 2012 MEAP 3-9 Item Bank including both operational and field-test items and then Assessment and Evaluation Services (AES) replicated those item statistics according to the specified layout by BAA in order to verify those item stats. Analysis was completed in two phases for the Fall 2012 MEAP administration. One is the operational items and the other is the field-test items in order to select future operational items. The item stats verified by AES were delivered to the BAA via the secure website in the ascribed layout and then confirmed by BAA. Finally, completed item stats were updated to the BAA item bank system by BAA psychometrician.

## CHAPTER 9: RELIABILITY

Reliability refers to the consistency and the precision of the scores obtained from a test as a measuring instrument. Of particular concern is the consistency with which the test measures the same individual on different occasions or with equivalent sets of items. Stated in another way, measures of reliability make it possible to estimate what proportion of total score variance is associated with random errors of measurement. Evidence of reliability, summarized typically by a reliability coefficient, is usually reported as an estimate of internal consistency or temporal stability. The extent to which errors of measurement are present in obtained scores is shown by the standard error of measurement (SEM).

### 9.1. Internal Consistency, Empirical IRT Reliability Estimates, and Conditional Standard Error of Measurement

In the following sections, estimates of reliability under classical test theory, the conditional standard error of measurement constructed under item response theory, and the use of standard error of measurement will be discussed in detail.

#### 9.1.1. Internal Consistency

Internal consistency is a measure of how well a collection of items work together to measure the construct. Typically, this index is computed as coefficient alpha or Cronbach's alpha. Coefficient alpha is a more general version of the common Kuder-Richardson reliability coefficients and can accommodate both dichotomous and polytomous items. The formula for Cronbach's alpha is:

$$\alpha = \left( \frac{k}{k-1} \right) \left( 1 - \frac{\sum (SD_i)^2}{(SD_x)^2} \right),$$

where

$k$  = number of items,

$SD_i$  = standard deviation of item  $i$ ,

$SD_x$  = standard deviation of the total scores.

*Appendix O* shows the coefficient alpha for each grade and subject. Other than the overall sample, the alphas are also computed by form, gender, ethnicity, social economic status (SES), and limited English proficiency (LEP).

Table 9.1.1.1 summarizes the coefficient alphas across forms in low and high ranges and median values for the overall population. Mathematics median IRT reliabilities are between .90 and .93. Reading median IRT reliabilities are in the range of .83 to .86, Science and social studies median reliabilities are in the 80's and Writing median reliabilities are .86 and .85 for grades 4 and 7, respectively. In conclusion, the MEAP assessments have a satisfactory degree of internal consistency as indicated by coefficient alphas.

By examining tables in *Appendix O*, similar conclusions also can be drawn from the comparisons between subgroups of gender, ethnicity, SES, and LEP. There is no obvious evidence that the alphas for subgroups, i.e., male versus female, are different. It is safe to conclude that the MEAP assessments are equally reliable for different subgroups.

**Table 9.1.1.1**  
**Summary Statistics of Coefficient Alphas**  
**Across Subjects and Grades**

<b>Subject</b>	<b>Grade</b>	<b>Low</b>	<b>High</b>	<b>Median</b>
Mathematics	3	0.91	0.91	0.91
Mathematics	4	0.91	0.92	0.91
Mathematics	5	0.91	0.92	0.92
Mathematics	6	0.91	0.92	0.92
Mathematics	7	0.92	0.93	0.93
Mathematics	8	0.90	0.90	0.90
Reading	3	0.85	0.85	0.85
Reading	4	0.85	0.86	0.85
Reading	5	0.86	0.87	0.86
Reading	6	0.84	0.85	0.85
Reading	7	0.86	0.86	0.86
Reading	8	0.82	0.83	0.83
Science	5	0.85	0.87	0.86
Science	8	0.84	0.88	0.87
Social Studies	6	0.81	0.87	0.84
Social Studies	9	0.87	0.88	0.87
Writing	4	0.86	0.86	0.86
Writing	7	0.85	0.86	0.85

### 9.1.2. Empirical IRT Reliability

In IRT, the precision of a test is shown by standard error of theta associated with each theta estimate and the test information function, which is also conditional on theta. However, by using the theta and standard error estimates from IRT and the definition of reliability from classical test theory (CTT), it is possible to derive an empirical IRT reliability from the data produced by IRT analysis. The variance of the theta estimates for the group can be treated as observed score variance. From the standard error of theta estimates, a pooled error variance can be created as an estimate of error variance. An empirical IRT reliability can then be derived from the formula:

$$\text{Empirical IRT Reliability} = [\text{Var}(\text{theta}) - \text{Var}(\text{error})] / \text{Var}(\text{theta}).$$

*Appendix P* shows the empirical IRT reliability for each grade and subject. The empirical IRT reliability coefficients are computed by form, gender, ethnicity, social economic status, and LEP group.

Table 9.1.2.1 summarizes the empirical IRT reliabilities across forms in low and high ranges and median values for the overall population. Mathematics median IRT reliabilities are between .88 and .92. Reading median IRT reliabilities are in the range of .78 to .90, Science median IRT reliabilities are between .84 and .88. Social studies median reliabilities are between .84 and .87, and Writing median reliabilities are close to .90. Reliabilities computed from empirical IRT approach yield similar results as alpha reliabilities.

In comparing IRT reliabilities between subgroups such as male vs. female, there is also no obvious evidence of differences in IRT reliabilities. It can be concluded that the MEAP are expected to be equally reliable for different subgroups such as gender, SES, ethnicity, and LEP.

**Table 9.1.2.1**  
**Summary Statistics of Empirical IRT Reliabilities**  
**Across Subjects and Grades**

Subject	Grade	Low	High	Median
Mathematics	3	0.88	0.89	0.88
Mathematics	4	0.89	0.91	0.90
Mathematics	5	0.90	0.91	0.90
Mathematics	6	0.91	0.91	0.91
Mathematics	7	0.91	0.92	0.92
Mathematics	8	0.90	0.90	0.90
Reading	3	0.79	0.80	0.80
Reading	4	0.82	0.83	0.83
Reading	5	0.82	0.83	0.83
Reading	6	0.81	0.82	0.81
Reading	7	0.81	0.83	0.82
Reading	8	0.78	0.79	0.79
Science	5	0.84	0.86	0.85
Science	8	0.84	0.88	0.86
Social Studies	6	0.84	0.86	0.85
Social Studies	9	0.86	0.87	0.86
Writing	4	0.90	0.91	0.91
Writing	7	0.89	0.89	0.89

### 9.1.3. Conditional Standard Error of Measurement

As previously described, information function is used in item response theory to gauge the precision of a test. Item information function for one parameter Rasch model is:

$$I_i(\theta) = P_i(\theta)[1 - P_i(\theta)],$$

where  $P_i(\theta)$  = the probability of correct answer in the Rasch model.

It is clear that the item information is maximized when ability parameter equals item difficulty (that is, when  $P_i(\theta) = .5$ ). When an item's difficulty is well matched to an examinee's ability, the item will

yield the most information about the latent trait. In IRT, test information function is the sum of item information function in the test, and the conditional standard error of measurement for a given ability estimate is the reciprocal of the square root of the test information function.

Test information function and conditional standard error of measurement are presented in plots in chapter 7, section 7.1.2.4., by form, grade, and subject. The plots show that as the test information function increases, the conditional standard error of measurement decreases and vice versa.

#### **9.1.4. Use of the Standard Error of Measurement**

In CTT, standard error of measurement (SEM) is calculated by the following formula:

$$SEM = \sigma_x \sqrt{1 - \rho_{xx'}}$$

where

$\sigma_x$  = standard deviation of the total test (standard deviation of the raw scores),

$\rho_{xx'}$  = reliability estimate for the test.

In contrast to the conditional SEM produced under IRT, the SEM produced under CTT is considered to be less precise because it is likely to be influenced by the characteristics of the sample (sample dependent). In addition, there is only one SEM applied equally to every examinee in the sample regardless of the score level of the examinee.

The MEAP adopts conditional SEM produced under IRT in the interpretation and reporting of individual student scores. Conditional SEM is considered to be more accurate than the traditional SEM produced under CTT, both theoretically and practically. The conditional SEM based on test information function is a sample-free estimate because it is not influenced by the sample characteristics. It is influenced only by the model and item parameters.

### **9.2. Alternative Forms Reliability Estimates**

At this time, no information regarding alternative forms reliability estimates is available for the technical report because no student takes more than one form of the MEAP during any test administration.

### **9.3. Score Reliability for the Written Composition and the Constructed-Response Items**

#### **9.3.1. Reader Agreement**

To ensure that all written compositions and the constructed-response items generated for MEAP are reliably scored, Measurement Incorporated (MI) uses several measures to gauge score reliability. One measure of reliability has been expressed in terms of reader agreement as obtained from the required second reading of a percentage of student responses. These data are monitored on a daily basis by the Scoring Monitor during the scoring process. Reader agreement data show the percent of perfect agreement of each reader against all other readers. For grades 3 - 8, 20 percent of all responses are given a second reading.

Reader agreement data do not provide a mechanism for monitoring drift from established criteria by all readers at a particular grade level. Thus, an additional set of data, known as validity scoring, is collected daily to check for reader drift and reader consistency in scoring to the established criteria.

When MI team leaders identify ideal student responses, they route these to the scoring directors for preview. Scoring directors review the responses and choose appropriate papers for validity scoring. Validity responses are usually solid score point responses. The scoring directors confirm the true score and enter the response for validity scoring. Readers score a validity packet every week for reading and writing. Validity scoring is blind; because image based scoring is seamless, scorers do not know when they are scoring a validity packet. Results of validity scoring are regularly analyzed by MI scoring directors, and appropriate measures are initiated as needed, including the retraining or releasing of scorers.

### 9.3.2. Score Appeals

A district may appeal the score assigned to any student's composition about which a question has been raised. In these instances, MI provides an individual analysis of the composition in question.

## 9.4. Estimates of Classification Accuracy

Every test administration will result in some error in classifying examinees. Several elements of test construction and guidelines around setting cut scores can assist in minimizing these errors. However, it is important to investigate the expected level of misclassification before approval of the final cut-scores. BAA conducts analysis of the classification accuracy of assessments based on the cut-scores recommended by college readiness study.

Under the IRT model, for a given ability score  $\theta$ , the observed score  $\hat{\theta}$  is expected to be normally distributed with a mean of  $\theta$  and a standard deviation of  $SE(\theta)$ . The expected proportion of examinees with true scores in any particular level  $k$  is:

$$\Pr(Level_k) = \sum_{\theta=c}^d \left( \phi \left( \frac{b-\theta}{SE(\theta)} \right) - \phi \left( \frac{a-\theta}{SE(\theta)} \right) \right) \varphi \left( \frac{\theta-\mu}{\sigma} \right),$$

where  $a$  and  $b$  are scale points representing the score boundaries (cut-scores) for levels,  $d$  and  $c$  are the scale points representing score boundaries for persons in levels,  $\phi$  is the cumulative distribution function of the achievement level boundary, and  $\varphi$  is the normal density associated with the true score (Rudner, 2004).

To compute expected classification accuracy, the proportions are computed for all cells of a  $K$  by  $K$  classification table. The sum of the diagonal entries represents the overall classification accuracy for the test.

### 9.4.1. Statewide Classification Accuracy

Tables in *Appendix R* show the results of statewide classification accuracy by grade and subject. In the

classification tables, the rows represent the theoretical true percentages of examinees in each achievement level while the columns represent the observed percentages. The diagonal entries represent the agreement between true and expected percentages of classified examinees. The sum of the diagonal representing total classification accuracy is presented in the bottom of the tables and summarized in Table 9.4.1.1. As Table 9.4.1.1 shows, the statewide classification accuracy rates range from 79.9% to 82.8% for Mathematics; 74.9% to 81.7% for Reading; 81.5% to 83.7% for Writing, 79.5% to 80.7% for Science and 79.3% to 79.5% for Social Studies.

**Table 9.4.1.1**  
**Summary of Statewide Classification Accuracy**

<b>Grade</b>	<b>Mathematics</b>	<b>Reading</b>	<b>Writing</b>	<b>Science</b>	<b>Social Studies</b>
3	81.3	77.1			
4	80.3	81.7	83.7		
5	82.2	78.7		79.5	
6	81.6	74.9			79.3
7	82.8	75.7	81.5		
8	79.9	75.5		80.7	
9					79.5

## **CHAPTER 10:**

### **VALIDITY**

Validity refers to the extent to which a test measures what it is intended to measure and how well it does so. As stated in the *Standards for Educational and Psychological Testing* (1999), validity refers to the “degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests.” This statement shows that test validation is an ongoing process, which begins the moment that work on a test begins and continues throughout the life of the test. Validity is the process of continually accumulating and reviewing evidence from various resources to refine the utility of a test for making recommended interpretations consistent with the intended uses and interpretations of the test scores. Thus, this chapter considered all types of evidence about validity issues.

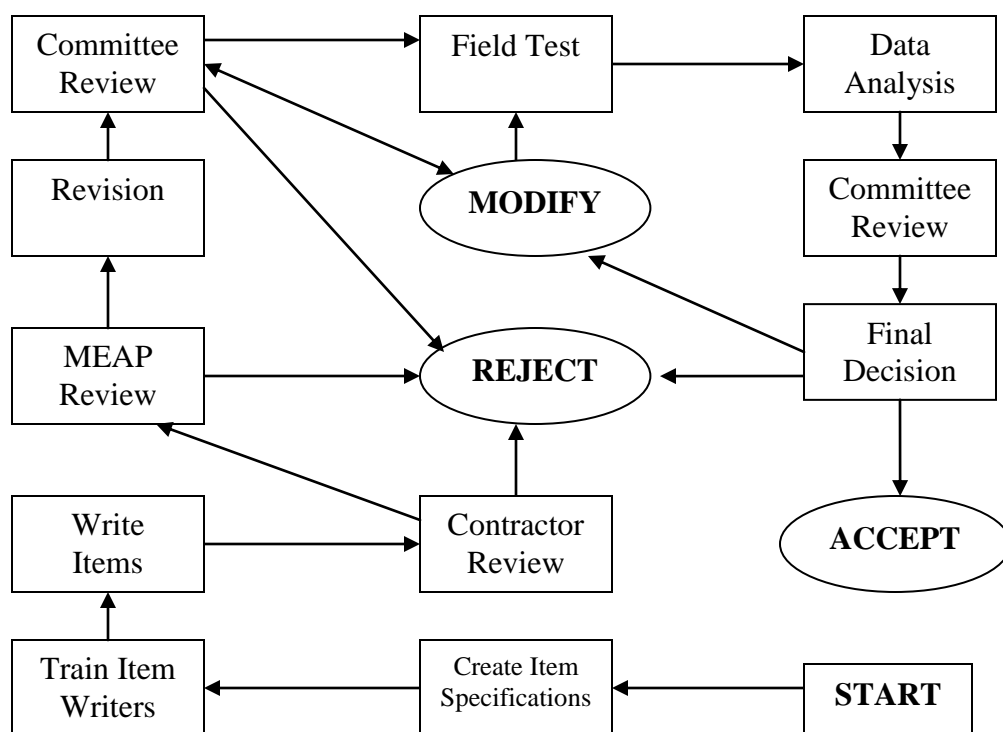
#### **10.1. Content and Curricular Validity**

Content validity involves essentially the systematic examination of the test content to determine whether it covers the curricular standards to be measured. As stated in Chapter 1, the MEAP assessments are developed to measure what Michigan educators believe all students should know and be able to achieve in the content areas. Assessment results paint a picture of how Michigan students and schools are doing when compared with standards established by the State Board of Education. The MEAP is based on an extensive definition of the content the test is intended to assess and its match to the content standards. Therefore, the MEAP assessments are content-based and aligned directly to the statewide content standards.

##### **10.1.1. Relation to Statewide Content Standards**

From the inception of the MEAP, a committee of educators, item development experts, assessment experts, and Bureau of Assessment and Accountability (BAA) staff met annually to review new and field-tested items. The BAA has established a sequential review process, as illustrated in Figure 10.1.1. This process provides many opportunities for these professionals to offer suggestions for improving or eliminating items and to offer insights into the interpretation of the statewide content standards for the MEAP. These review committees participate in this process to ensure test content validity of the MEAP.

In addition to providing information on the difficulty, appropriateness, and fairness of these items, committee members provide a needed check on the alignment between the items and the content standards they are intended to measure. When items are judged to be relevant (i.e., representative of the content defined by the standards), this provides evidence to support the validity of inferences made (regarding knowledge of this content) with MEAP results. When items are judged to be inappropriate for any reason, the committee can either suggest revisions (e.g., reclassification or rewording) or elect to eliminate the item from the field-test item pool. Items that are approved by the content review committee are later embedded in live MEAP forms to allow for the collection of performance data. In essence, these committees review and verify the alignment of the test items with the objectives and measurement specifications to ensure that the items measure appropriate content. The nature and specificity of these review procedures provide strong evidence for the content validity of the MEAP.



**Figure 10.1.1. MEAP Item Development/Review Cycle**

#### 10.1.1.1. MEAP Alignment Studies

Three alignment studies (ELA, mathematics, and science) were conducted on September 21, 22, and 23, 2005 in Lansing, Michigan.

##### *ELA*

For ELA, twelve reviewers, including language arts content experts, district language arts supervisors, and language arts teachers, met to analyze the agreement between the state's language arts standards and assessments for grades 3 through 8. Eight reviewers were from Michigan, and four were experts with experience from other states.

The alignment between the assessments and the language arts standards at each grade was acceptable. The over-emphasis on one or two reading objectives on the assessment is not a critical alignment issue, since all of the other alignment criteria were fully met. The alignment between the assessments and the writing standard at each grade needs slight improvement. One to three more objectives for each grade need to have at least one corresponding item for the assessments to fully meet the Range-of-Knowledge Correspondence criterion. Reviewers were very consistent in assigning items to standards, but showed less consistency in assigning items to specific grade-level expectations. This implies some overlap in content covered by the grade-level expectations, or lack of clarity in the written statements. Because reviewers found it difficult to distinguish among many of the objectives, this lowered the reviewer

agreement on the precise objective measured by an item. The reviewers observed that the coverage of content on the assessments improved over the grades. Reviewers indicated there were some very challenging items on the grade 7 and grade 8 assessments.

The complete report on ELA alignment analysis can be found in *Appendix S*. The report consists of a description of the four criteria used to judge the alignment between Michigan Language Arts Academic Content Standards and one assessment for each grade, and includes a description of the alignment criteria used and a complete presentation of the findings.

## ***Mathematics***

For mathematics, thirteen reviewers, including mathematics content experts, district mathematics supervisors, mathematics teachers, and a mathematics education professor, met to analyze the agreement between the state's mathematics standards and Michigan Educational Assessment Program assessments for six grades. Ten reviewers were from Michigan, and 3 were experts brought in from other states. Twelve to thirteen reviewers analyzed grades 3, 4, and 5 assessments, while 6 or 7 reviewers analyzed grades 6, 7, and 8 assessments. Because of time constraints the reviewers were divided into two groups to analyze the assessments for the higher grades. All of the reviewers participated in analyzing the depth-of-knowledge levels of the standards.

Overall, the alignment between the mathematics assessments and standards at five of the six grades is reasonable. The grade 6 assessment was fully aligned. Full alignment between the assessments at grades 4, 5, 7, and 8 and the previous grade standards could be achieved by replacing one item (grades 4, 5, and 8) or three items (grade 7) on each assessment. Full alignment for the grade 3 assessment and grade 2 standards would require replacing six items with items that measure content related to data and probability. Reviewers did have some problems coding items to specific standards because limits imposed on number size and type of number in the grade-level expectations did not fully coincide with the numbers used in the items. As a consequence, reviewers coded a relatively large number of items to the goal or standard rather than to specific grade level expectations. The lack of exact fit between an assessment item and a grade-level expectation could be due to the grade-level expectations being overly restrictive, or to test blueprint specifications that did not attend to the stated limits. Although, reviewers did code a relative high number of items to the goal or standard, this was not such a serious issue as to consider the assessments and standards not aligned. A complete report of the mathematics alignment study is included in *Appendix T*. The report consists of a description of the four criteria used to judge the alignment between Michigan Mathematics Academic Content Standards and one assessment for each grade, and includes the alignment criteria used and a complete description of the findings.

## ***Science***

For science, nine reviewers, including science content experts, the state science coordinator, district science supervisors, and science teachers met to analyze the agreement between the state's science standards and the Michigan Educational Assessment Program assessments for grades 5 and 8. Five reviewers were from Michigan, and four were experts brought in from other states. Nine reviewers analyzed four of the six assessments, while four analyzed the grade 8 2004 assessment and six reviewers analyzed the grade 8 2005 assessment. Because of time constraints, the reviewers were divided into two groups to analyze these two assessments. All of the reviewers participated in analyzing the depth-of-knowledge levels of the standards.

The Michigan science standards and assessments for grades 5 and 8 lack full alignment because one standard is not assessed. Reviewers at most only coded three items to Standard II (*Reflecting on Scientific Knowledge*) on any of the six forms analyzed. On most forms, reviewers found no items that they judged to correspond to objectives under this standard. Many of the objectives under this standard seek to have students develop an awareness of the nature of science or an application of science, which are more difficult to measure on an on-demand assessment. Considering the assessments and the other four standards for both grade levels, the alignment is reasonable, with only a few changes needed to achieve full alignment. If the three forms at each grade level are considered in aggregate, then the combined test is fully aligned with the four standards.

If each assessment form is thought of as a separate assessment, then only a few changes to each form are needed to achieve acceptable alignment between the assessment and the science standards. Each grade 5 form would need to have only one or two items replaced or added to meet the minimal acceptable levels on all four alignment criteria. The grade 8 forms would require from three to five additional items, or replaced items, to achieve an acceptable alignment on the four alignment criteria; in each case, for each of the six forms, it would be possible to retain the total number of items and have full alignment if existing items were replaced by new items. A full report of science alignment study is provided in *Appendix U*. The report consists of a description of the four criteria used to judge the alignment between Michigan Science Academic Content Standards for grades 5 and 8 and three assessments for each grade, and includes the alignment criteria used as well as a complete description of the findings.

#### **10.1.2. Educator Input**

Michigan educators provide valued input on the MEAP content and the match between the items and the statewide content standards. In addition, many current and former Michigan educators and some educators from other states work as independent contractors to write items specifically to measure the objectives and specifications of the content standards for the MEAP. Using a varied source of item writers provides a system of checks and balances for item development and review that reduces single source bias. Because many people with various backgrounds write the items, it is less likely that items will suffer from a bias that might occur if items were written by a single author. This direct input from educators offers evidence regarding the content validity of the MEAP.

#### **10.1.3. Test Developer Input**

The staff at the BAA provide a long history of test development experience, including content-related expertise. The input and review by these assessment experts provide further support of the item being an accurate measurement of the intended objective. These reviews offer additional evidence of the content validation of the MEAP.

#### **10.1.4. Evidence of Content Validity**

As stated above, expert judgments from educators, test developers, and assessment specialists provide support for the alignment of the MEAP assessments with the statewide content standards. In addition, since expert teachers in the content areas were involved in establishing the content standards, the judgments of these expert teachers in the review process provides a measure of content validity. A

match between the content standards and the components of the MEAP provides evidence that the assessment does indeed measure the content standards. The MEAP test blueprint discussed in sections 2.1 and

2.2 shows the number of assessment components, tasks, or items matching each content standard, providing documentation of the content validity of the assessment.

## **10.2. Criterion and Construct Validity**

### **10.2.1. Criterion Validity**

Criterion validity refers to the degree to which a test correlates with other external outcome criteria. Criterion validity addresses how accurately criterion performance can be predicted from test scores. The key to criterion-related evidence is the degree of relationship between the assessment and the outcome criterion. To ensure a good relationship between the assessment and the criterion, the criterion should be relevant to the assessment and reliable.

As previously stated, test validation is an ongoing process that continues throughout the life of the test. At this point, the MEAP does not have any criterion-related validity to be reported. However, for the purposes of this technical report and with respect to the MEAP, additional criterion-related validity evidence will be collected and reported in an ongoing manner. These data are most likely to come from districts conducting program evaluation research, universities as well as special interest groups researching topics of local interest, as well as the data collection efforts of the BAA.

### 10.2.2. Construct Validity

The construct validity of a test refers to the extent to which the test is said to measure a theoretical construct or trait. A structural equation modeling study conducted by Dr. Joseph Martineau of BAA is summarized in this section as evidence of construct validity.

In order to demonstrate that the assessment scores are strongly related to variables they should be related to (e.g. prior achievement) and not strongly related to variables they should not be related to (e.g. student demographics), it is necessary to have at least two years of data for each student. Table 10.2.2.1 demonstrates where it is possible to conduct a valid analysis with two years of data for the same children.

**Table 10.2.2.1 Student Achievement Data Available in BAA Database**

Assessment	Subject	Cycle	Grade								HS
			3	4	5	6	7	8	...		
MEAP	Mathematics	Winter 2005		x				x		x	
		Fall 2005	x	x	x	x	x	x		x	
	ELA	Winter 2005		x			x			x	
		Fall 2005	x	x	x	x	x	x		x	
	Science	Winter 2005			x			x		x	
		Fall 2005			x			x		x	
MI-Access	Math & ELA	Winter 2005									
		Fall 2005	x	x	x	x	x	x		x	

Winter 2005 was the first assessment cycle in Michigan with strong student identifiers (or unique identifying codes). Therefore, it is the first year of data for which strong longitudinal links can be created. There is only one additional assessment cycle (Fall 2005) with data available for matching. Therefore, the Winter 2005 cycle of data is the only one that can validly serve to create a baseline for two cycles of data for each student. As shown in Table 10.2.2.1, there must be data in the winter 2005 cycle and data in the next grade for the Fall 2005 cycle to be able to match student data across multiple cycles. This is only available in MEAP for two grades in English Language Arts (ELA), for one grade in Mathematics, and not available at all for MI-Access because MI-Access was a new assessment in Fall 2005.

Therefore, this analysis is conducted only for two grades in MEAP ELA, and for one grade in MEAP Mathematics. It was determined not to conduct such analyses for any other grades because it is important to have prior achievement in the model. Where prior achievement is not contained in the model, the analysis of relationships as intended is misleading.

To analyze the relationships as intended, the structural equation model shown in Figure 10.2.2.1 was performed for each of the three possible analyses. The structural equation model as shown is a saturated model because all possible relationships among variables are included in the model. This saturation does not affect the results being interpreted for the purpose of this report. There are several pieces of the model. First, all possible covariances between exogenous variables (first order predictors) are included in the model to accommodate the relationships between those demographic variables. Second, each

exogenous variable is designated as a predictor of both achievement scores. Third, the prior achievement is designated as a predictor of later achievement.

By designing the model in this way, it is possible to examine the direct effect of demographics on current achievement in light of the effect of prior achievement on current achievement. This can be compared to the overall effect of demographics on current achievement to determine how much of the relationship between demographics and current achievement can be explained by the relationship between prior achievement and current achievement.

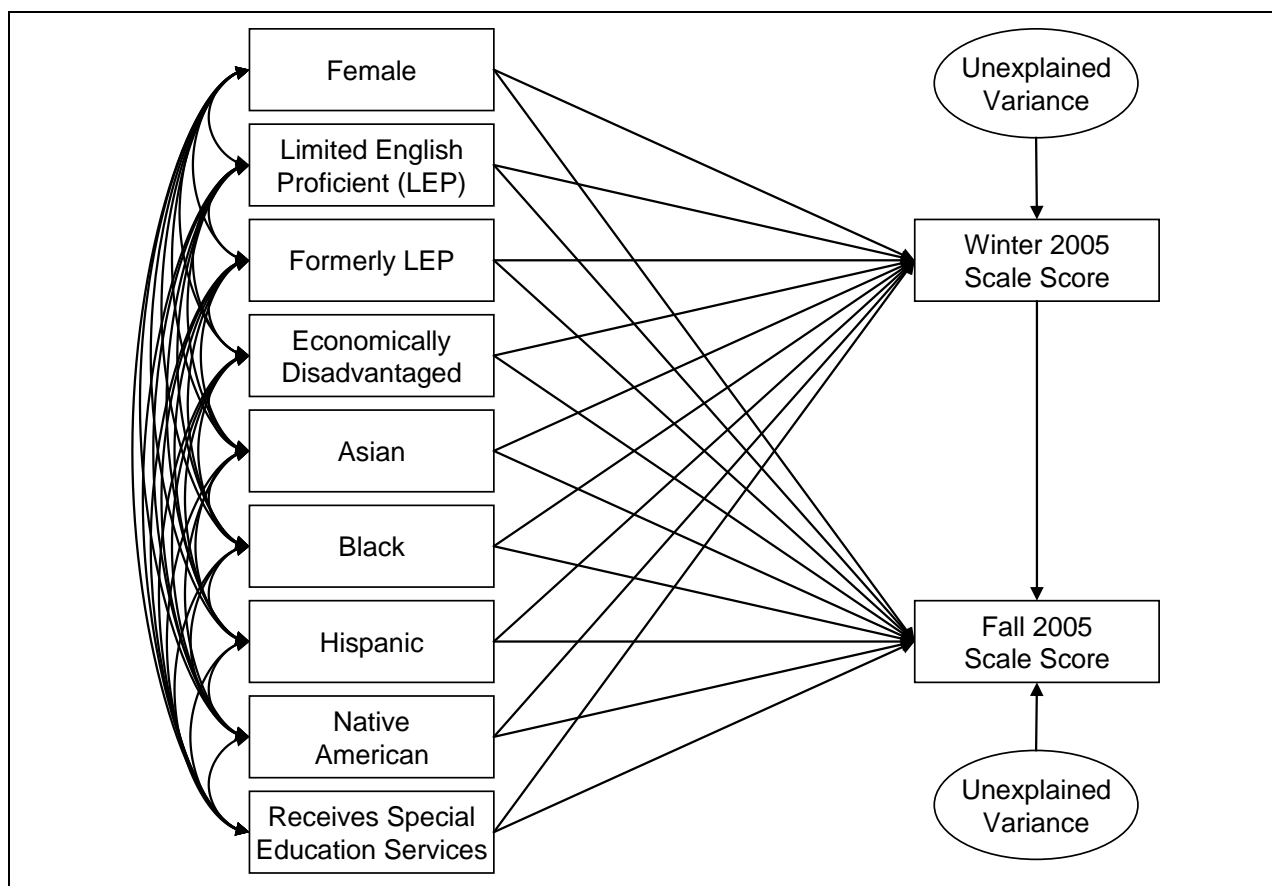


Figure 10.2.2.1. Basic Structural Equation Model

### Grade 4 Mathematics Results

The effect sizes of the total and direct effects of the demographics on current achievement are presented in Table 10.2.2.2. Table 10.2.2.2 also gives the effect size reduction (in terms of percentages) in the relationship between demographics and current achievement that can be attributed to prior achievement.

As shown by Table 10.2.2.2, accounting for prior achievement reduces the effect of demographic variables on current achievement by anywhere from 40 to 71 percent. In addition, the effect size of prior achievement on current achievement from this model is 0.68. In other words, the effect of prior achievement on current achievement is from

5.3 to 136.0 times as large as the effects of demographic variables on current achievement. In addition, the magnitude of the effect size of prior achievement is a large effect size (0.60), while the magnitude of direct effect sizes of demographic variables are from negligible (0.005) to small (0.128).

**Table 10.2.2.2**  
**Grade 4 Mathematics effect sizes of demographic variables on current achievement**

Demographic Variable	Effect Size		Reduction
	Total	Direct	
Female	-0.051	-0.025	51%
Limited English Proficient (LEP)	-0.051	-0.016	69%
Formerly LEP	0.010	0.006	40%
Economically Disadvantaged	-0.198	-0.075	62%
Multiracial	-0.017	-0.005	71%
Hispanic	-0.054	-0.020	63%
Black	-0.255	-0.128	50%
Asian	0.083	0.035	58%
Native American	-0.025	-0.009	64%
Receives Special Education Services	-0.208	-0.076	63%

In addition, while the data needed to assess teacher and/or school effects on student achievement (theoretically the most important predictor) is not available, percentage of variation remaining unexplained in current achievement from this model (39%) is high enough to allow for strong sensitivity to instruction even when accounting for prior achievement and demographics.

#### ***Grade 4 ELA Results***

The effect sizes of the total and direct effects of the demographics on current achievement are presented in Table 10.2.2.3. Table 10.2.2.3 also gives the effect size reduction (in terms of percentages) in the relationship between demographics and current achievement that can be attributed to prior achievement.

**Table 10.2.2.3**  
**Grade 4 ELA effect sizes of demographic variables on current achievement**

Demographic Variable	Effect Size		Reduction
	Total	Direct	
Female	0.105	0.023	78%
Limited English Proficient (LEP)	-0.070	-0.024	66%
Formerly LEP	0.010	0.005	50%
Economically Disadvantaged	-0.205	-0.096	53%
Multiracial	-0.008	-0.001	88%
Hispanic	-0.037	-0.016	57%
Black	-0.196	-0.091	54%
Asian	0.052	0.028	46%
Native American	-0.023	-0.009	61%
Receives Special Education Services	-0.249	-0.107	57%

As shown by Table 10.2.2.3, accounting for prior achievement reduces the effect of demographic variables on current achievement by anywhere from 46 to 88 percent. In addition, the effect size of prior achievement on current achievement from this model is 0.61. In other words, the effect of prior achievement on current achievement is from 5.7 to 610.0 times as large as the effects of demographic variables on current achievement. In addition, the magnitude of the effect size of prior achievement is a large effect size (0.61), while the magnitude of direct effect sizes of demographic variables are negligible (0.001) to small (0.107).

In addition, while the data needed to assess teacher and/or school effects on student achievement (theoretically the most important predictor) is not available, percentage of variation remaining unexplained in current achievement from this model (49%) is high enough to allow for strong sensitivity to instruction even when accounting for prior achievement and demographics.

### ***Grade 7 ELA Results***

The effect sizes of the total and direct effects of the demographics on current achievement are presented in Table 10.2.2.4. Table 10.2.2.4 also gives the effect size reduction (in terms of percentages) in the relationship between demographics and current achievement that can be attributed to prior achievement.

**Table 10.2.2.4. Grade 7 ELA effect sizes of demographic variables on current achievement**

Demographic Variable	Effect Size		Reduction
	Total	Direct	
Female	0.142	0.053	63%
Limited English Proficient (LEP)	-0.068	-0.021	69%
Formerly LEP	-0.007	-0.003	57%
Economically Disadvantaged	-0.184	-0.066	64%
Multiracial	-0.011	-0.006	45%
Hispanic	-0.042	-0.017	60%
Black	-0.221	-0.071	68%
Asian	0.061	0.034	44%
Native American	-0.029	-0.012	59%
Receives Special Education Services	-0.303	-0.118	61%

As shown by Table 10.2.2.4, accounting for prior achievement reduces the effect of demographic variables on current achievement by anywhere from 44 to 69 percent. In addition, the effect size of prior achievement on current achievement from this model is 0.67. In other words, the effect of prior achievement on current achievement is from 3.7 to 223.3 times as large as the effects of demographic variables on current achievement. In addition, the magnitude of the effect size of prior achievement is a large effect size (0.67), while the magnitude of direct effect sizes of demographic variables are to negligible (0.003) to small (0.118).

In addition, while the data needed to assess teacher and/or school effects on student achievement (theoretically the most important predictor) is not available, percentage of variation remaining unexplained in current achievement from this model (40%) is high enough to allow for strong sensitivity to instruction even when accounting for prior achievement and demographics.

### 10.3. Validity Evidence for Different Student Populations

The primary validity evidence of the MEAP lies in the content being measured. Since the test measures the statewide content standards, which are required to be taught to all students, the test is expected to be valid for use with all subpopulations of students. Because the MEAP measures what is required to be taught to all students and is given under the same standardized conditions to all students, the tests have the same validity for all students.

Moreover, great effort has been made to ensure that the MEAP items are fair and representative of the content domain expressed in the content standards. Additionally, much scrutiny is applied to the items and their possible effect on minority or other subpopulations making up the population of Michigan. Every effort is made to eliminate items that may have ethnic or cultural biases.

#### 10.3.1. Differential Item Functioning (DIF) Analyses

Two types of differential item functioning (DIF) are applied to the MEAP program in terms of editorial review and statistical DIF analyses.

### 10.3.1.1. Editorial Bias Review

In a typical editorial review, items are examined for three categories of bias: gender, ethnicity/class, and geographic region. The review process flags potentially biased items and codes them as follows:

**Status:** Are the members of a particular group shown in situations that do not involve authority or leadership?

**Stereotype:** Are the members of a particular group portrayed as uniformly having certain aptitudes, interests, occupations, or personality characteristics?

**Familiarity:** Is there greater opportunity on the part of one group to be acquainted with the vocabulary? Is there a greater chance that one group will have experienced the situation or have become acquainted with the processes presented by an item?

**Offensive Choice of Words:** Has a demeaning label been applied or has a male term been used where a neutral term could be substituted?

**Other:** Are there any other indications of bias?

Any potentially biased item is then edited to remove bias if possible. If that is not possible, the item is eliminated.

### 10.3.1.2. Statistical DIF Analyses

DIF statistics are used to identify items on which members of a focal group have different probability of getting the items correct from members of a reference group after they have been matched by means of the ability level on the test. In the MEAP DIF analyses, total raw score on the core items is used as ability-matching variable. Two comparisons are made for each item:

- Males (M) versus females (F)
- White (W) versus Black (B)

For the MC items, the Mantel-Haenszel Delta DIF statistics are computed (Dorans and Holland, 1992) to classify test items in three levels of DIF for each comparison: negligible DIF (A), moderate DIF (B), and large DIF (C). An item is flagged if it exhibits B or C category of DIF using the following rules derived from National Assessment of Educational Progress (NAEP) guidelines (Allen, Carlson, & Zalanak 1999):

Rule 1: MH *delta* (MHD) not significantly different from 0 (based on  $\alpha = .05$ ) or  $|MHD| < 1.0$  are classified as A. [Note: The MHD is the ETS delta scale for item difficulty, where the natural logarithm of the common odds ratio is multiplied by  $-(4/1.7)$ ]

Rule 2: MHD significantly different from 0 and  $\{|MHD| \geq 1.0 \text{ and } < 1.5\}$  or MHD not significantly different from 0 and  $|MHD| \geq 1.0$  are classified as B.

Rule 3:  $|MHD| \geq 1.5$  and significantly different from 0 are classified as C.

The effect size of the standardized mean difference (SMD) is used to flag DIF for the CR items. The SMD reflects the size of the differences in performance on CR items between student groups matched on the total score. The SMD is the difference between the unweighted item mean of the focal group and the weighted item mean of the reference group. The weights applied to the reference group are applied so that the weighted number of reference group students is the same as in the focal group (within the

same ability group). The SMD is divided by the total group item standard deviation to get a measure of the effect size for the SMD. The SMD effect size allows each item to be placed into one of three categories: negligible DIF (AA), moderate DIF (BB), or large DIF (CC).

Rule 1: If the probability is  $>.05$ , classify the item as AA.

Otherwise:

Rule 2: If  $|ES|$  is  $\leq .17$ , classify as AA.

Rule 3: If  $|ES|$  is  $>.17$  but  $\leq .25$ , classify as BB.

Rule 4: If  $|ES|$  is  $>.25$ , classify as CC.

For both MC and CR items the favored group is indicated if an item was flagged.

### **10.3.1.3. DIF Statistics for Each Item**

Every operational item in the Fall 2012 testing cycle had been field-tested. DIF analyses, extensive item analyses, and data review were performed to eliminate any possible bad and/or biased items. Data for embedded field-test items and operational items collected in the Fall 2012 testing cycle were analyzed again. In addition to the DIF analyses described above, detailed extensive item analyses were performed to establish complete item bank statistics. BAA conducted item and data review processes for the field test items as described in section 2.6. BAA also conducted item selection and test form construction for Fall 2012 testing cycle. For the operational items, item statistics (including DIF statistics) are reported in *Appendix H*.

### **10.3.2. Performance of Different Student Populations**

As previously described, great care has been taken in the process of test development to ensure that the items are fair and representative of the content domain expressed in the content standards. Additionally, every effort is made to eliminate items that may have gender, ethnic, or cultural biases through DIF analyses and qualitative content review. Even with all the effort to make the assessments fair, it is expected to see performance of various student populations being different. Tables in *Appendix V* show the student performance of MEAP assessment breakdown by gender, ethnicity, socioeconomic status, and English proficiency. Figures in *Appendix W* depict the scale score distributions with the same breakdown.

In addition, DIF analyses are also conducted on accommodated forms that have sufficient numbers of students for DIF analyses. Table 10.3.2.1 summarizes results of DIF analyses on male versus female and white versus black. In Table 10.3.2.1, numbers of items classified as A/AA (negligible DIF), B/BB (moderate DIF), and C/CC (large DIF) are tabulated. Overall, there are very low percentages of items that are classified as B/BB or C/CC.

**Table 10.3.2.1**  
**Summary of DIF analyses for MEAP Operational Items**

Subject	Grade	Number of Unique Items Across Forms	Male vs Female			White vs Black			With Accommodation vs Without Accommodation		
			A/AA	B/BB	C/CC	A/AA	B/BB	C/CC	A/AA	B/BB	C/CC
Mathematics	3	53	52	1		50	3		51	1	1
	4	59	59			53	6		58	1	
	5	54	52	2		52	2		53	2	
	6	60	56	4		60			60		
	7	62	57	5		59	3		62		
Reading	8	49	47	2		49			48	1	
	3	31	30		1	31			31		
	4	31	30	1		31			30	1	
	5	31	28	3		29	1	1	31		
	6	31	30	1		30	1		31		
Science	7	31	29	1	1	29	1	1	31		
	8	31	31			30	1		31		
	5	110	109	1		107	2	1	110		
	8	133	127	6		130	3		131	2	
	6	45	45			44	1		44		
Social Studies	9	44	43	1		44			44		
Writing	4	25	25			24	1		24	1	
	7	25	25			22	2	1	25		

A/AA : negligible DIF

B/BB : moderate DIF

C/CC : large DIF

#### 10.4 Validity Evidence for Accommodation Form (Person-Fit Analysis)

Additional study was implemented to check whether or not a high-stakes state assessment supports the assumption of measurement invariance/scale comparability by comparing non-accommodated forms with accommodated forms, and non-translated forms with translated forms for the science contents area. Based on the person-fit analysis, there was no evidence to suggest a grow violation of the measurement invariance assumption. The misfit ratios of the accommodated form were similar with the non-accommodated form, those of the translated form were also similar with the non-translated form in the science test. As a consequence, this study provided additional validity evidence that the inferences made based on  $\hat{\theta}$  and any subsequent linear transformations of  $\hat{\theta}$  are comparable across forms and accommodations. That is, the meaning of the scores at any point along the underlying ability continuum, as measured by the various forms of the assessments, are comparable and equally valid. The completed study was reported in *Appendix Y*.

10.5 Validity Evidence for Mode Comparability (Online vs. Paper-Pencil Tests)

Mode comparability study was implemented for Social Studies Grade 6 and 9. As a result, this study demonstrated that no special attention is necessary to offer simultaneously both online-based test (OBT) and paper-based test (PBT) assessments. With the propensity score matching method, the result corroborated that the scale scores implemented by OBT are comparable with those implemented by PBT in the first year transitioning into the computer technology for statewide assessment. The completed study was reported in *Appendix Z*. Furthermore, a post survey for students who took OBT showed that no more students felt uncomfortable with use of a computer. In a survey question of “how would you like to take/administer the MEAP test in the future”, about 70% preferred OBT mode, 10% preferred PBT mode, and the 20% preferred either way. Thus, student’s capability of a computer use in the statewide assessment would not affect their test scores at all.

10.6 MEAP Math Rescoring Issue

As was indicated in Chapter 2, there was an error in the identification of the content standards that have should have appeared on the MEAP Mathematics assessments that resulted in several of the operational items needing to be dropped from the initial scoring and analyses. In Math Grade 4, 7, and 8, this meant that the original tests which were supposed to be 59, 62, and 49 items ended up being 57, 60, and 43 items and had some shifts in the proportional distribution of items across the content domains. This section provides some comparisons of the original and rescored assessment results to show the impact that dropping the inappropriate items had in comparison to using the items that were originally supposed to appear on the assessment.

Table 10.6.1 shows the IRT based and Cronbach’s Coefficient Alpha reliability estimates for original and rescored versions of the assessment. One can see that the reliability estimates are lower for the rescored version of the Grade 7 Mathematics assessment. However, the drops in reliability in Grade 4 and 8 appear to be fairly same. The drop in reliability was 0.01 for Cronbach’s Coefficient Alpha for the grade 7 Mathematics assessment. This suggests that there were only small changes in the precision of the assessment for the original and rescored versions.

Table 10.6.1  
Reliability Estimates for Original and Rescored Versions

Grade	IRT based		Cronbach's Alpha	
	Original	Rescore	Original	Rescore
4	0.90	0.90	0.91	0.91
7	0.92	0.92	0.93	0.92
8	0.90	0.89	0.90	0.90

Table 10.6.2 shows the scale score summary statistics for all students for original and rescored versions of the assessments. One can see that the scale score means and scale score standard deviations are very similar to each other. The differences in scale score means ranged from -0.13 to 0.13 scale score points and the differences in scale score standard deviations ranged from 0.10 to 0.59. Again, the rescore

seemed to have a very small impact on student scores at the aggregate level.

Table 10.6.2  
Summary Statistics for Original and Rescored Versions

Grade	Original				Rescore				Difference (Rescore-Original)			
	Scale Score		SE		Scale Score		SE		Scale Score		SE	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
4	431.40	24.94	7.55	2.16	431.53	25.71	7.87	2.48	.13	2.42	.32	.79
7	725.16	29.72	8.01	2.65	725.12	29.81	8.12	2.70	-.05	1.68	.10	.34
8	820.22	29.12	8.89	2.19	820.10	29.54	9.48	2.48	-.13	4.00	.59	.82

Table 10.6.3 shows the results for performance level classifications for the original and rescored versions of the assessments. One can see some small, but important differences in the percent of students in each performance level across grades for the original and rescored versions. Most notably, the differences in percent in the Advanced level was 2.03 in Grade 4, in the Proficient level was -2.52 in Grade 8, and in the Not Proficient level was -1.96 in Grade 8. Since these numbers are small, they suggest that roughly these aggregate changes in percentages of students changed from a proficient to not proficient performance were trivial.

Table 10.6.3  
Percent of Students by Proficiency Level for Original and Rescored

Grade	Advanced			Proficient			Partially Proficient			Not proficient		
	Rescore	Original	Differ	Rescore	Original	Differ	Rescore	Original	Differ	Rescore	Original	Differ
4	8.27	6.24	2.03	37.92	38.82	-.90	14.52	18.24	-3.72	38.39	36.70	1.69
7	6.02	5.73	.27	32.28	32.83	-.55	24.11	23.39	-.72	37.58	38.04	-.46
8	8.19	8.71	-.52	26.57	24.05	-2.52	26.34	26.38	-.04	38.88	40.84	-1.96

Another way of looking at the performance level changes is to look at the students who had exactly the same classification versus students that had different classifications for the original and rescored versions. These results are shown in Table 10.6.4. One can see that the percentage of students with the same classification on the original and rescore versions were 94.45% for Grade 4, 96.98% for Grade 7, and 91.12% for Grade 8. The percentage of students who saw their performance level on the rescored version increase were 3.51% for Grade 4, 1.75% for Grade 7, and 6.10% for Grade 8, while the percentage of students who saw their performance level on the rescored version decrease were 2.04% for Grade 4, 1.26% for Grade 7, and 2.66% for Grade 8. These number do not exactly sum to zero or to the values in Table 4 because it was possible for tradeoffs to occur, such as some students having their score go up and other students their score down, and in addition it was possible for a student to see their score increase or decrease by more than one performance level. For some students, whether the classification was based on the original or rescored versions of the assessment definitely did make a difference. This does present some challenges related to using the scores for accountability purposes.

Table 10.6.4  
Percent of Students with Same versus  
Different Performance Level Classifications  
for Original and Rescored Versions

Grade	Same	Higher	Lower
4	94.45	3.51	2.04
7	96.98	1.75	1.26
8	91.22	6.10	2.66

## **CHAPTER 11:**

### **ACCOUNTABILITY USES OF ASSESSMENT DATA**

The major policy-based uses of assessment data from MEAP and MI-Access are for public reporting and school accountability decisions.

#### **11.1. Legislative Grounding**

- Throughout 2011-2012, the Michigan Department of Education (MDE) worked with local education stakeholders across the state, as well as with the United States Department of Education (USED) to develop a request for flexibility from certain requirements of the Elementary and Secondary Education Act (ESEA), also known as "No Child Left Behind." The flexibility requested includes waivers of 11 specific provisions of this federal law, including the requirement that all schools meet the 100% student proficiency targets by 2014. Michigan's work in implementing career- and college-ready expectations for all students; developing differentiated recognition, accountability, and support for districts and schools; and supporting effective instruction and leadership will create the context in which this flexibility may be successfully implemented for the benefit of Michigan's students and education community.

On July 19, 2012, the MDE received notification from USED that Michigan's ESEA Flexibility Request was approved. Michigan statute (section 1280 of the Revised School Code) requires the State Board of Education to accredit public elementary and secondary schools. The State Board approved *Education YES – A Yardstick for Excellent Schools!* in 2002, and accepted the report of the Accreditation Advisory Committee in 2003.

The Michigan School Accountability Scorecards combine student assessment data with graduation or attendance rates as well as information on compliance with state and federal laws. The Scorecard is a diagnostic tool that gives schools, districts, parents, and the public an easy way to see a school's or district's strengths and weaknesses.

The Michigan School Accountability Scorecards are a replacement to the Michigan School Report Cards that were required under the No Child Left Behind (NCLB) Act to report Adequate Yearly Progress (AYP). Michigan received an Elementary and Secondary Education Act (ESEA) Flexibility Waiver from the U.S. Department of Education in July 2012 that allows the use of the Scorecards in place of the former AYP Report Cards.

*Education YES!* uses several components that are interlinked to present a complete picture of performance at the school level. *Education YES!* is a broad set of measures that looks at school performance and looks at student achievement in multiple ways. Measures of student achievement in Michigan's school accreditation system include:

- Achievement status to measure how well a school is doing in educating its students.
- Achievement change to measure whether student achievement is improving or declining.

In addition, the Indicators of School Performance measure investments that schools are making in improved student achievement, based on indicators that come from research and best practice.

## **11.2. Procedures for Using Assessment Data for Accountability**

Targets for participation, proficiency, and graduation or attendance must be met for the school or district as a whole and for any valid subgroup. There are 12 potential subgroups for a school and 13 potential subgroups for a district. The minimum size for a subgroup is almost always 30 students. The one exception to the minimum size is for the Bottom 30% subgroup. The minimum size required for the Bottom 30% subgroup is 9 students. The “All Students” group will display even if the entire school or district has fewer than 30 students. The subgroups include:

- Major Racial/Ethnic Groups
  - Black or African American
  - American Indian or Alaska Native
  - Asian American
  - Native Hawaiian or other Pacific Islander
  - Hispanic or Latino
  - White
  - Multiracial
- Students with Disabilities
- Limited English Proficient
- Economically Disadvantaged
- Bottom 30% (achievement only)
- Shared Educational Entity (SEE) for district AYP only

Michigan’s minimum subgroup size is 30 students. For a district or school that enrolls more than 3,000 students, the minimum subgroup size will be 1% of enrollment, up to a maximum subgroup size of 200 students. An accountability determination will be made for all subgroups of 200 or more students.

It is the policy of the Michigan State Board of Education that all students participate in the state assessment program. The student’s status in terms of enrollment for a full academic year is not relevant to whether the student should be assessed. The federal No Child Left Behind Act requires that at least 95% of enrolled students be assessed. The number of students to be assessed is determined from the Michigan Student Data System (MSDS) collected by the Center for Educational Performance and Information (CEPI). This is taken from the Fall (September) collection for grades 3-9 and from the Spring (February) collection for high schools.

Proficiency targets are unique to each school and district. Targets are set at the school and district level in each content area. This means that any subgroup present in the school or district must meet the school or district’s proficiency target. All schools and districts are expected to reach 85% proficiency in all content areas by the end of the 2021-22 school year.

Proficiency targets are based on the school or district’s full academic year percent proficient in 2011-12. Proficient students are those who attain a Performance Level 1 or 2 on the MEAP, MME, MEAP-Access, or MI-Access. This initial proficiency rate is called the base year percent proficient. The targets

for each successive year are incremented equally over ten years by taking the difference between 85% and the base year percent proficient. Targets are calculated for each subject assessed in a school or district.

Individual school and district proficiency targets can be found here:

[http://www.michigan.gov/documents/mde/Michigan\\_Proficiency\\_Targets\\_413516\\_7.xls](http://www.michigan.gov/documents/mde/Michigan_Proficiency_Targets_413516_7.xls)

Because the decisions made based upon accountability classifications are such high-stakes decisions for individual schools, it is important to account for error in order to be accurate in classifying schools as meeting or not meeting their accountability targets. Uncertainty in scores has an impact on classifying students as proficient, and uncertainty in classifying students as proficient has an impact on calculating accountability. For this reason, measurement error needs to be taken into account in calculating accountability. Measurement error can cause two types of errors in calculating accountability: false positives (mistakenly identifying schools as meeting targets) and false negatives (mistakenly identifying schools as not meeting targets).

Students with scale scores within two conditional standard errors of measurement of the proficient cut score are considered provisionally proficient for accountability.

*Education YES!* uses several components that are interlinked to present a complete picture of performance at the school level. *Education YES!* is a broad set of measures that looks at school performance and looks at student achievement in multiple ways. Measures of student achievement in Michigan's school accreditation system include:

- Achievement status to measure how well a school is doing in educating its students.
- Achievement change to measure whether student achievement is improving or declining.

In addition, the Indicators of School Performance measure investments that schools are making in improved student achievement, based on indicators that come from research and best practice. Scores on all three components of *Education YES!* have been converted to a common 100 point scale where: 90-100 A; 80-89 B; 70-79 C; 60-69 D; and 50-59 F. Grades of D and F are not used for the school's composite grade, where the labels D/Alert and Unaccredited are used.

### **Achievement Status**

Achievement status is measured in reading and mathematics at the elementary level. It includes science and social studies at the middle school and high school levels. Achievement Status uses up to three years of comparable data from the Michigan Educational Assessment Program (MEAP).

The method of computing achievement status uses students' scale scores on the Michigan Educational Assessment Program, as weighted by the performance level or category (1,2,3, or 4) assigned to each student's score. Scale score values at the chance level are substituted for values below the chance level because values below that point do not have valid information about the student's performance. A template is provided so that a school can paste in MEAP data to see how the values are derived. The weighted index is computed by following these steps:

1. Multiply each student's scale score by the performance level (i.e., 540\*2);
2. Sum of the resulting values resulting in the sum of the index values;
3. Sum of the performance levels or weights;
4. Divide the sum of the index values by the sum of the weights.

The intent of the weighted index is to encourage schools to place priority on improving the achievement of students that attain the lowest scores on the MEAP assessments.

Cut scores for the score ranges in achievement status were set by representative panels that assigned grades to selected schools. The cut scores were reviewed by the Accreditation Advisory Committee and approved by the State Board of Education. The Accreditation Advisory Committee, a group of five national experts, was appointed by the State Board of Education to advise the Board on the implementation of the *Education YES!* school accreditation.

### **Achievement Change**

Achievement change uses up to five years of comparable MEAP data to determine if student achievement in a school is improving at a rate fast enough to attain the goal of 85% proficiency in school year 2021-22, as required by the ESEA Flexibility Waiver. The change grade is derived from the average of up to three calculations of improvement rates (slopes) using the school's MEAP data. Scores from MEAP assessments that are not comparable will not be placed on the same trend line.

The Achievement Change component of *Education YES!* was originally proposed to recognize improvement on the part of schools with low status scores. The Accreditation Advisory Committee recommended a policy-based approach to measuring achievement change. Achievement change uses up to five years of comparable MEAP data to determine if student achievement in a school is improving at a rate fast enough to attain the goal of 85% proficient by school year 2021-22, as required by Michigan's approved ESEA Flexibility Waiver. The change grade is derived from the average of three calculated slopes using the school's MEAP and MI-Access data. Scores from MEAP assessments that are not comparable will not be placed on the same slope line. Achievement Change is based on the goal of 85% percent proficient in 2021-22, as set in the ESEA Flexibility Waiver. Achievement Change is computed by dividing the computed slope by the target slope, determining the percent of the target that the school has attained.

Multiple linear regression was used to predict each school's 2012-13 score based on the school's scores from 2011-12, and 2010-11. A prediction was made for each content area and grade level that was tested in previous years. The prediction was compared to the school's actual 2012-13 percent proficient. The Difference is computed as the (Actual – Predicted). The school's status score for each content area and grade range is adjusted as follows:

- Schools where the actual score exceeds the prediction plus 1.5 times the standard error of the estimate had a 15 point adjustment added to the achievement score for that content area;
- Schools where the actual score exceeds the prediction plus the standard error of the estimate had a 10 point adjustment added to the achievement score for that content area;
- Schools where the actual score is less than the prediction minus 1.5 times the standard error of the estimate had a 15 point deduction applied to the achievement score for that content area; and

- Schools where the actual score is less than the prediction minus the standard error of the estimate had a 10 point deduction applied to the achievement score for that content area.

The Achievement Change adjustment was calculated only if there are at least 10 students tested each year (2010-11, 2011-12 and 2012-13) in the content area and grade level.

The composite school grade is derived from the school scores and letter grades and the school's Accountability Scorecard status. The weighting of the components of *Education YES!* in the composite grade has been as follows:

***Education YES! Composite Score  
Weighting***

<b>Component</b>	<b>Point Value</b>
School Performance Indicators	33
Achievement Status	34
Achievement Change	33
Total	100

The scores for each content area will be averaged to calculate an achievement score and grade for each school. An achievement score for each content area has been computed by averaging the Status and Change (or adjusted Change) scores for a content area. A preliminary aggregate achievement score is derived by averaging the scores from each content area. The preliminary aggregate achievement score is weighted 67% and the School Self-Assessment (Indicator score) is weighted 33% in calculating the preliminary score and grade for a school.

In 2004-05, the State Board of Education approved a change to the *Education YES!* policy so that the school's indicator score cannot improve the school's composite score and grade by more than one letter grade more than the school's achievement grade. This means that a school that receives an "F" for achievement can receive a composite grade no higher than "D/Alert."

After the computation of a school's composite grade for achievement described above a final "filter" will be applied, consisting of the question of whether or not a school or district met or did not meet its accountability targets. The answer to this question is an additional determining factor for a school's final composite grade on the report card. A school that does not make its accountability targets shall not be given a grade of "A." A school that makes its accountability targets shall not be listed as unaccredited. A school's composite school grade will be used to prioritize assistance to underperforming schools and to prioritize interventions to improve student achievement.

***Unified Accountability for Michigan Schools***

<b><i>Education YES!</i></b> <b>Composite Score</b>	90-100	B (iv)	A
	80-89	B (iv)	B (iv)
	70-79	C (iii)	C (iii)
	60-69	D/Alert (ii)	C (iii)
	50-59	Unaccredited (i)	D/Alert (ii)
		<b><i>Did Not Make Targets</i></b>	<b><i>Makes Targets</i></b>

(i) – (iv) Priorities for Assistance and Intervention

Schools that are labeled “A”, “B”, “C”, or “D / Alert” will be accredited. Schools that receive an “A” will be summary accredited. Schools that receive a “B”, “C”, or “D/Alert” will be in interim status. Unaccredited schools will also be labeled as such. Summary accreditation, interim status, and unaccredited are labels from Section 1280 of the Revised School Code.

Results of accountability analyses for 2012-13 are reported in next section. Results of accountability analyses for 2013-14 will be available in August, 2014.

### **11.3. Results of Accountability Analyses**

3397 School Accountability Scorecards

- 2011-12: 3411 School Report Cards

2886 (85%) Schools made accountability targets (orange or higher):

- 2011-12: 2726 (79.9%) Schools made AYP

511 (15%) Schools did not meet accountability targets (red):

- 2011-12: 602 (17.6%) Schools did not make AYP

873 District Accountability Scorecards

- 2011-12: 543 District Report cards

181 (20.7%) Districts did not meet accountability targets (red)

- 2011-12: 259 (47.7%) Districts did not make AYP

692 (79.3%) Districts met accountability targets (orange or higher)

- 2011-12: 284 (52.2%) Districts made AYP

## REFERENCE

- Allen, N. L., Carlson, J. E., and Zalanak, C. A. (1999). *The NAEP 1996 Technical Report*. Washington, DC: National Center for Education Statistics.
- American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for the educational and psychological testing*. Washington, DC: Author.
- Dorans, N. J., and Holland, P. W. (1992). DIF Detection and Description: Mantel-Haenszel and Standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 35–66). Hillsdale, NJ: Erlbaum.
- Kolen, M. J., & Brennan, R. L. (2004). *Test Equating: Methods and Practices*. (2nd ed.). New York: Springer-Verlag.
- Martineau, J. A. (2007, March). Designing a valid and transparent progress-based value-added accountability model. *Paper presented at the Annual Conference of the American Educational Research Association (AERA), Chicago, IL.*
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- Masters, G. N. & Evans, J. (1986). Banking non-dichotomously scored items. *Applied Psychological Measurement*, 10(4), pp. 355-367.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.
- Rudner, L. M. (2004, April). *Expected Classification Accuracy*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.
- Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14(2), pp. 97-116.
- Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis: Rasch Measurement*. Chicago: MESA Press.
- Wright, B. D., and Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.
- Wyse, A. E., Zeng, J., & Martineau, J. A. (2011). A graphical transition table for communicating status and growth. *Practical Assessment, Research and Evaluation*, 16(11). Retrieved from <http://pareonline.net/getvn.asp?v=16&n=11>.

## **LIST OF APPENDICES**

- Appendix A: Summary Statistics of Matrix Sampling
- Appendix B: Data Created for Field Test Items
- Appendix C: Computation of DIF Statistics
- Appendix D: Statistics Used on Item Labels
- Appendix E: New Developed Cut Scores
- Appendix F: Updated Revised Accommodation Summary Table
- Appendix G: Sample Output for Key Check
- Appendix H: Item Statistics
- Appendix I: Standard Setting Technical Report
- Appendix J: Classical Item Statistics
- Appendix K: Scale Score Statistics and Performance Level Percentage
- Appendix L: Histogram of Scale Score Distributions
- Appendix M: Summary of Item Response Theory Statistics
- Appendix N: Summary of Item Fit Statistics
- Appendix O: Alpha Reliabilities
- Appendix P: Empirical IRT Reliabilities
- Appendix Q: Rater Agreement
- Appendix R: Statewide Classification Accuracy
- Appendix S: ELA Alignment Study Technical Report
- Appendix T: Mathematics Alignment Study Technical Report
- Appendix U: Science Alignment Study Technical Report
- Appendix V: Performance of Different Student Populations
- Appendix W: Histogram of Scale Score Distributions by Student
- Appendix X: Independent Quality Assurance Review
- Appendix Y: Person-fit Analysis
- Appendix Z: Mode Comparability Study